Wykład dla studiów doktoranckich IMDiK PAN

Biostatystyka I

dr Anna Rajfura
Kat. Doświadczalnictwa i Bioinformatyki SGGW
anna_rajfura@sggw.pl

Program wykładu w skrócie

1. Hipotezy o normalności rozkładu.
   Testy: chi-kwadrat zgodności, Shapiro-Wilka, Kołmogorowa-Smirnowa z poprawką Lilliforsa.
2. Porównanie dwóch prób niezależnych.
   Testy: t-Studenta, U Manna-Whitneya, Walda-Wolfowitza.

Powtórzenie. Statystyczny opis danych

Dane liczbowe z pomiaru: $x_1, x_2, \ldots, x_n$ w badaniu pełnym; parametry: $\bar{x}$, s^2, s.

$\bar{x}$ – średnia arytmetyczna charakteryzuje średni poziom, wokół którego skupiają się wartości ze zbioru danych

s^2 – wariancja pokazuje rozrzut danych wokół średniej arytmetycznej

s – odchylenie standardowe pokazuje, o ile przeciętnie poszczególne wyniki różnią się od średniej, czyli pokazuje wielkość błędu pojedynczego pomiaru

Powtórzenie cd.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad s = \sqrt{s^2}$$

Według tych wzorów liczą funkcje wbudowane arkusza kalkulacyjnego:

=ŚREDNIA
=WARIANCJA.POPUL
=ODCH.STANDARD.POPUL

Czasami badacz chciałby mieć pogląd na wartość średnią czy wariancję pełnego zbioru wyników, choć nie ma możliwości zbadania wszystkich interesujących go jednostek statystycznych.

Powtórzenie. Wnioskowanie

losowanie

Populacja generalna → Próba

$x_1, x_2, \ldots, x_n$

cecha X w populacji

parametry populacyjne:

średnia μ

odch. standardowe σ

cecha X w próbie

parametry próby:

średnia $\overline{x}$

odch. standardowe s

wnioskowanie

Metody wnioskowania:

- **estymacja punktowa, przedziałowa**
- **testowanie hipotez**

Powtórzenie. Estymacja punktowa

Estymacja **punktowa** parametru populacyjnego polega na podaniu **jednej liczby** będącej oceną tego parametru. Tę liczbę oblicza się na podstawie próby (mówimy: parametr próby, a właściwie: statystyka próby).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

wariancja nieobciążona

$$s = \sqrt{s^2}$$

ze wzoru na wariancję nieobciążoną

Powtórzenie. Estymacja punktowa

Wartości obliczane na podstawie próby dają funkcje wbudowane arkusza kalkulacyjnego:

=ŚREDNIA
=WARIANCJA
=ODCH.STANDARDOWE

Także pakiet *STATISTICA* w opcji **Statystyki opisowe** oblicza parametry próby (według powyższych wzorów).

Powtórzenie. Wnioskowanie

Metody wnioskowania:

- **estymacja punktowa, przedziałowa** (parametrów populacyjnych)
- **testowanie hipotez** (o parametrach populacyjnych lub typie rozkładu cechy w populacji)

Niektóre z metod wymagają, aby próba pochodziła z rozkładu normalnego.

Testowanie hipotez statystycznych

1. Sformułowanie hipotezy statystycznej
2. Losowanie próby
3. Wybór testu (funkcji testowej) do zbadania hipotezy i ustalenie poziomu istotności α
4. Wyznaczenie wartości funkcji testowej dla próby
5. Sformułowanie wniosku odnośnie hipotezy

Testowanie hipotezy o normalności rozkładu – na przykładzie.

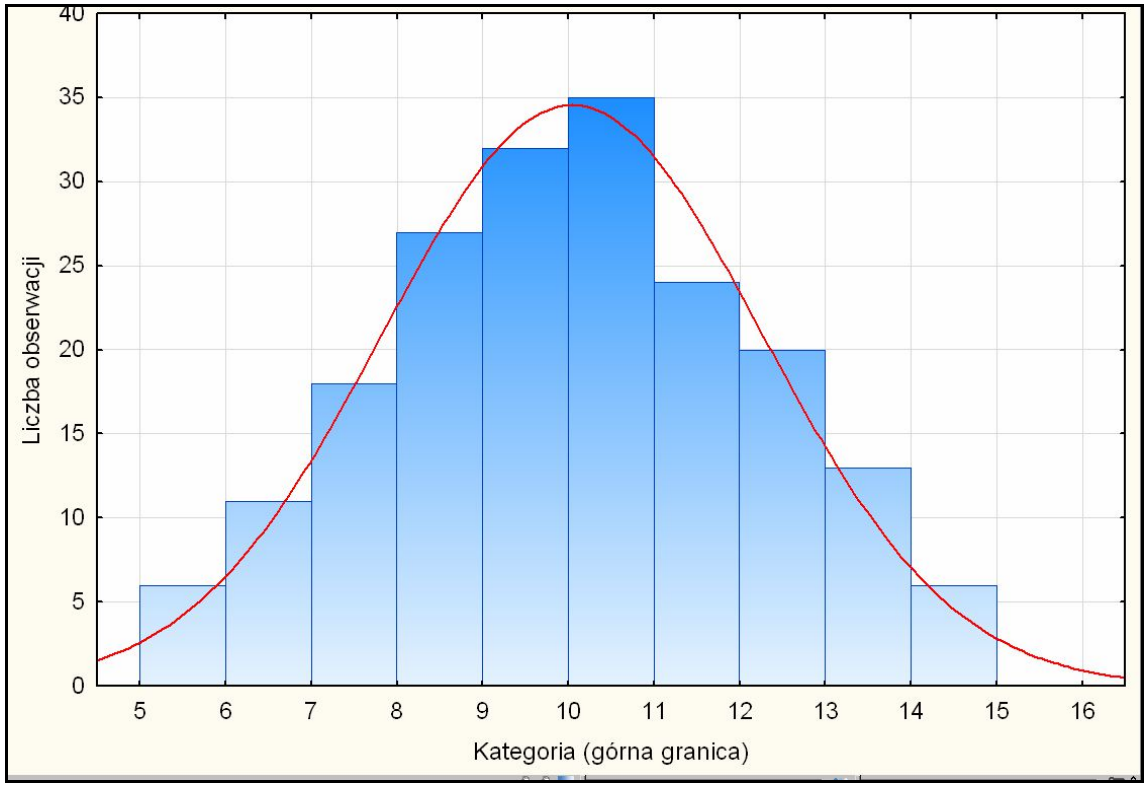
Badanie normalności rozkładu-przykład

Podczas eksperymentu oceniano wartość pewnego parametru biochemicznego (PB) w grupie 192 chorych. Wyniki zestawiono w szeregu rozdzielczym.

Zakres wartości	Liczebność
[5; 6]	6
(6; 7]	11
(7; 8]	18
(8; 9]	27
(9; 10]	32
(10; 11]	35
(11; 12]	24
(12; 13]	20
(13; 14]	13
(14; 15]	6
Razem	192

Przykład

Rysunek przedstawia histogram dla próby w porównaniu z przebiegiem krzywej normalnej.



Test chi-kwadrat zgodności

Populacja: chorzy na rozpatrywaną chorobę

Cecha X - wartość określonego parametru biochemicznego PB u chorych

Sprawdzimy, czy cecha X ma rozkład normalny.

Hipoteza zerowa, ozn.: H_0

H_0: cecha X ma rozkład normalny

Hipoteza alternatywna, ozn.: H_1

H_1: cecha X nie ma rozkładu normalnego

Test chi-kwadrat zgodności cd.

Próbę stanowią dane zestawione w szeregu rozdzielczym.

Zakres wartości	Liczebność
[5; 6]	6
(6; 7]	11
(7; 8]	18
(8; 9]	27
(9; 10]	32
(10; 11]	35
(11; 12]	24
(12; 13]	20
(13; 14]	13
(14; 15]	6
Razem	192

Test chi-kwadrat zgodności cd.

Wybieramy test chi-kwadrat, ozn. χ^2

Wzór funkcji testowej:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i^t\right)^2}{n_i^t}$$

gdzie:

k- liczba klas w szeregu rozdzielczym

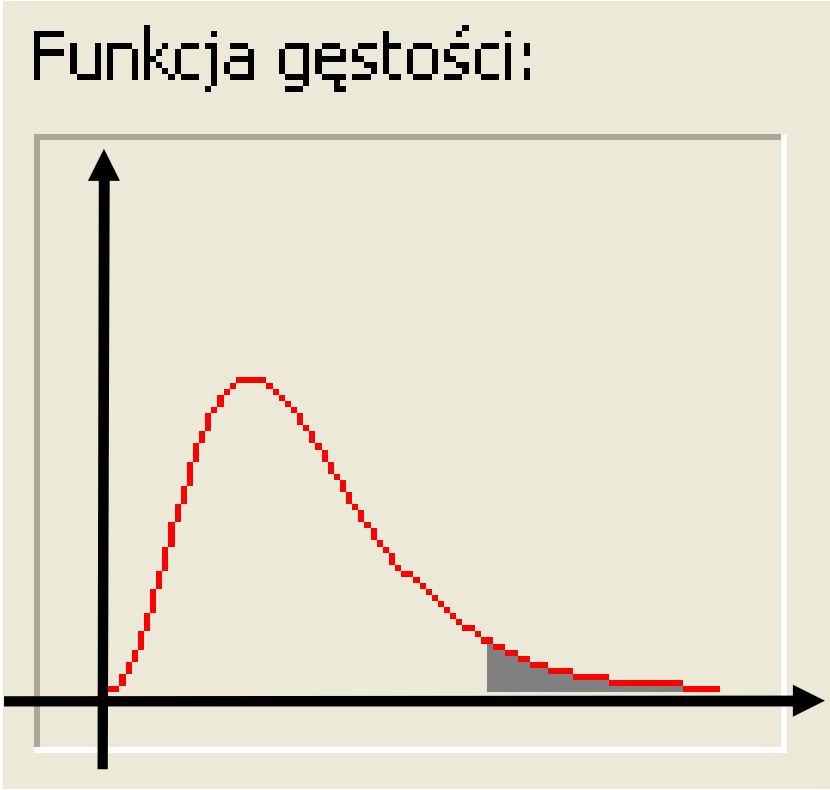
n_i – liczebność empiryczna w *i*-tej klasie

n_i^t – wartość p-stwa teoretycznego w *i*-tej klasie

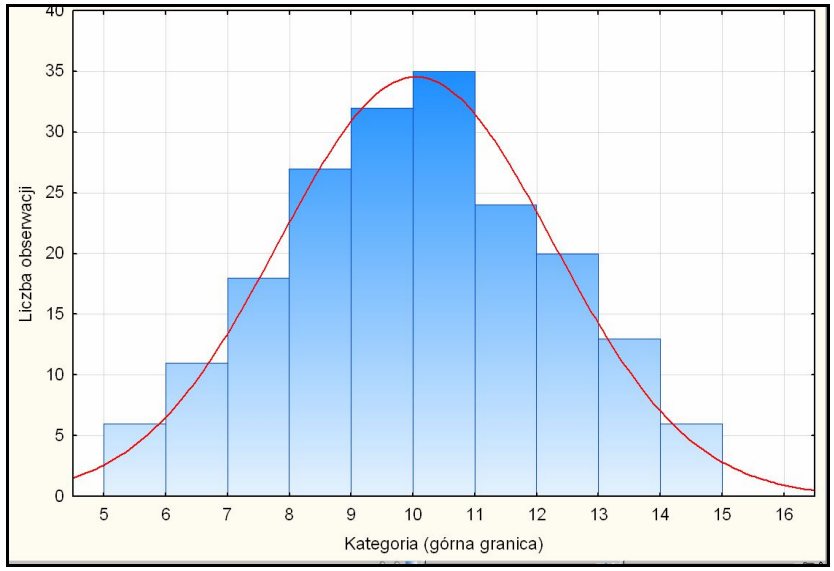
Karl Pearson podał w tym teście kryterium, na podstawie którego podejmujemy decyzję – jest nim różnica pól zakreślonych przez wykres rozkładu empirycznego i teoretycznego.

Test chi-kwadrat zgodności cd.

Funkcja testowa ma rozkład chi-kwadrat o liczbie stopni swobody $\nu = k\text{-}u\text{-}1$, gdzie k jest liczbą klas w szeregu rozdzielczym, u liczbą parametrów rozkładu szacowanych na podstawie próby.



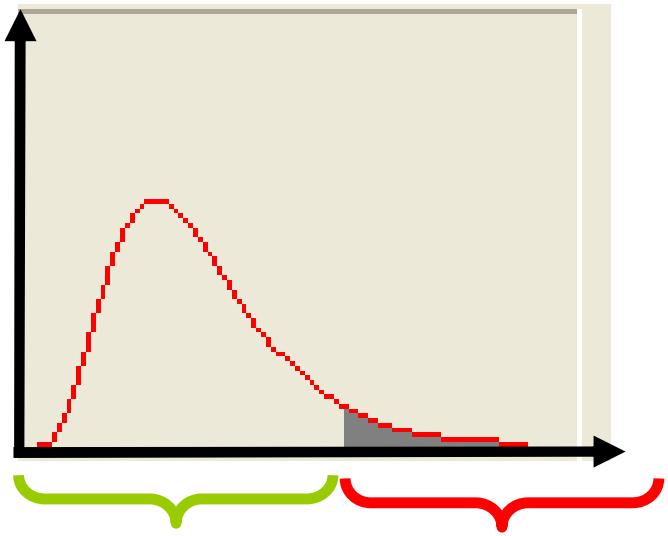
Test chi-kwadrat zgodności cd.



$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i - n_i^t\right)^2}{n_i^t}$$

Test chi-kwadrat zgodności cd.

W układzie współrzędnych na osi poziomej są możliwe wartości funkcji testowej.



Obszar dopuszczenia hipotezy

Obszar odrzucenia hipotezy

Test chi-kwadrat zgodności cd.

W układzie współrzędnych na osi poziomej są możliwe wartości funkcji testowej.

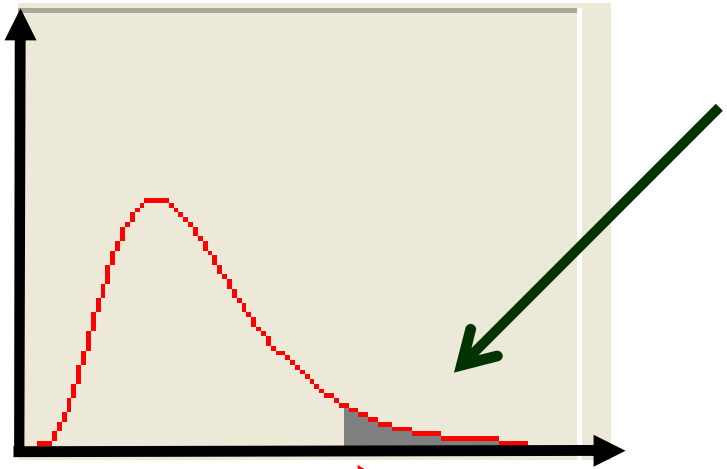
Pole wynosi α

Obszar odrzucenia hipotezy

Liczba α wyraża p-stwo popełnienia błędu odrzucenia hipotezy prawdziwej, α nazywamy **poziomem istotności**.

Test chi-kwadrat zgodności cd.

W przykładzie przyjmijmy poziom istotności α = 0,05.



Pole wynosi α=0,05

Wartość krytyczna funkcji testowej

Wartość krytyczną można odczytać z tablic statystycznych.

Test chi-kwadrat zgodności cd.

poziom istotności $\alpha = 0{,}05$

liczba stopni swobody $\nu = 8$

Wartości krytyczne rozkładu chi-kwadrat

$X \sim \chi^2_\nu$ - X zmienna losowa o rozkładzie chi-kwadrat z liczbą stopni swobody ν,
α - poziom istotności,
$\chi^2_{\alpha,\nu}$ - wartość krytyczna - liczba taka, że $P(X > \chi^2_{\alpha,\nu}) = \alpha$

ν \ a	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	$0{,}0^4393$	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2104	10,5965
3	0,0717	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449	12,8381
4	0,2070	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767	14,8602
5	0,4118	0,5543	0,8312	1,1455	1,6103	9,2363	11,0705	12,8325	15,0863	16,7496
6	0,6757	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119	18,5475
7	0,9893	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,3444	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902	21,9549
9	1,7349	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,1558	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093	25,1881
11	2,6032	3,0535	3,8157	4,5748	5,5778	17,2750	19,6752	21,9200	24,7250	26,7569
12	3,0738	3,5706	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	26,2170	28,2997
13	3,5650	4,1069	5,0087	5,8919	7,0415	19,8119	22,3620	24,7356	27,6882	29,8193
14	4,0747	4,6604	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	29,1412	31,3194

Test chi-kwadrat zgodności cd.

Wartość krytyczna funkcji testowej wynosi χ^2_{kryt}=15,5073.

Teraz obliczymy wartość funkcji testowej dla próby.

Test chi-kwadrat zgodności cd.

Zakres wartości	Liczebność empiryczna n_i	P-stwo pi
[5; 6]	6	
(6; 7]	11	
(7; 8]	18	
(8; 9]	27	
(9; 10]	32	
(10; 11]	35	
(11; 12]	24	
(12; 13]	20	
(13; 14]	13	
(14; 15]	6	
Razem	192	

Test chi-kwadrat zgodności cd.

Obliczanie p-stwa zdarzenia:

$$X \in \langle a; b)$$

z rozkładu normalnego

$$X \sim N(\mu;\ \sigma^2)$$

Można wykonać ze wzoru:

$$P\{X \in \langle a; b)\} = F_X(b) - F_X(a)$$

Po zestandaryzowaniu dystrybuanty, wartości można przeczytać z tablic statystycznych:

$$P\{X \in \langle a; b)\} = F_X(b) - F_X(a) = F_Z\left(\frac{b-\mu}{\sigma}\right) - F_Z\left(\frac{a-\mu}{\sigma}\right)$$

Test chi-kwadrat zgodności cd.

W omawianym przykładzie nie znamy parametrów rozkładu

$$X \sim N(?;\ ?)$$

zatem szacujemy je na podstawie próby:

$$\hat{\mu} = \overline{x} = 10{,}044, \quad \hat{\sigma}^2 = s^2 = 4{,}916$$

i dopiero obliczamy p-stwa potrzebne w szeregu rozdzielczym.

Uwaga. Trzeba było oszacować dwa parametry.

Test chi-kwadrat zgodności cd.

Zakres wartości	Liczebność empiryczna n_i	P-stwo pi	Liczebność teoretyczna $n_i^t=p_i \cdot 192$
[5; 6]	6	0,034	
(6; 7]	11	0,051	
(7; 8]	18	0,093	
(8; 9]	27	0,141	
(9; 10]	32	0,173	
(10; 11]	35	0,175	
(11; 12]	24	0,144	
(12; 13]	20	0,098	
(13; 14]	13	0,054	
(14; 15]	6	0,037	
Razem	192	1,000	

Test chi-kwadrat zgodności cd.

Zakres wartości	Liczebność empiryczna n_i	P-stwo pi	Liczebność teoretyczna $n_i^t=p_i \cdot 192$	Składnik funkcji testowej
[5; 6]	6	0,034	6,541	
(6; 7]	11	0,051	9,754	
(7; 8]	18	0,093	17,933	
(8; 9]	27	0,141	26,992	
(9; 10]	32	0,173	33,260	
(10; 11]	35	0,175	33,554	
(11; 12]	24	0,144	27,714	
(12; 13]	20	0,098	18,740	
(13; 14]	13	0,054	10,374	
(14; 15]	6	0,037	7,139	
Razem	192	1,000		

Test chi-kwadrat zgodności cd.

Zakres wartości	Liczebność empiryczna n_i	P-stwo pi	Liczebność teoretyczna $n_i^t=p_i \cdot 192$	Składnik funkcji testowej
[5; 6]	6	0,034	6,541	0,0448
(6; 7]	11	0,051	9,754	0,1591
(7; 8]	18	0,093	17,933	0,0003
(8; 9]	27	0,141	26,992	0,0000
(9; 10]	32	0,173	33,260	0,0477
(10; 11]	35	0,175	33,554	0,0623
(11; 12]	24	0,144	27,714	0,4976
(12; 13]	20	0,098	18,740	0,0847
(13; 14]	13	0,054	10,374	0,6647
(14; 15]	6	0,037	7,139	0,1816
Razem	192	1,000		**1,7429**

Test chi-kwadrat zgodności cd.

Wartość krytyczna funkcji testowej wynosi χ^2_{kryt}=15,5073.

Wartość funkcji testowej dla próby wynosi χ^2=1,74.

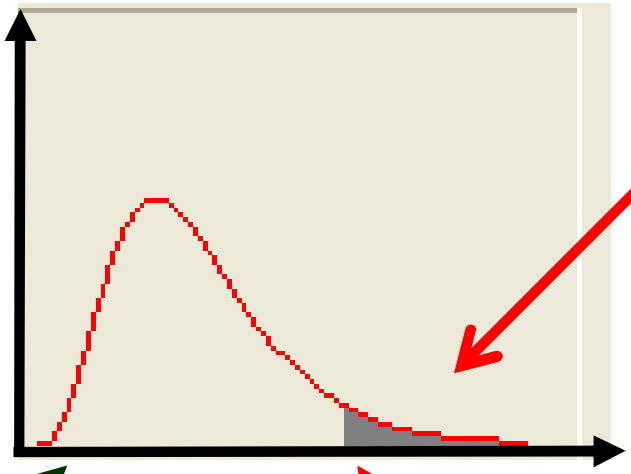
Wnioskowanie

Jeżeli $\chi^2 > \chi^2_{kryt}$, to hipotezę zerową odrzucamy,

w przeciwnym przypadku hipotezy nie można odrzucić.

W przykładzie: hipotezy o normalności rozkładu cechy w populacji nie odrzucamy.

Test chi-kwadrat zgodności cd.



Pole wynosi

$\alpha = 0{,}05$

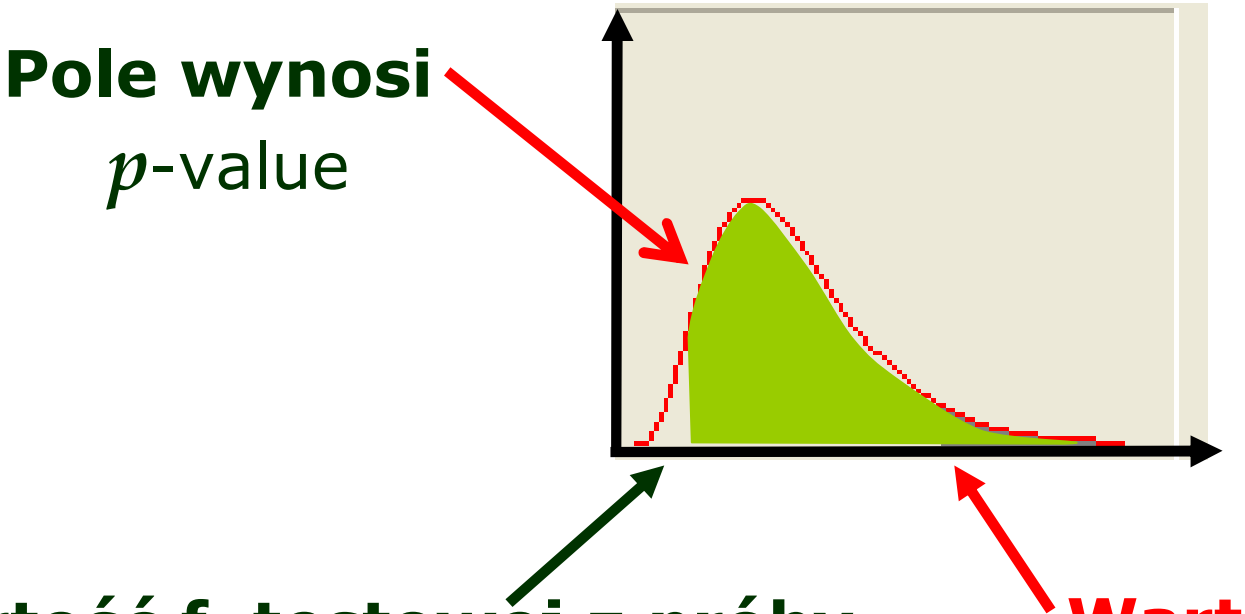
Wartość f. testowej z próby

1,74

Wartość krytyczna

15,5073

Test chi-kwadrat zgodności cd.



Wnioskowanie równoważne

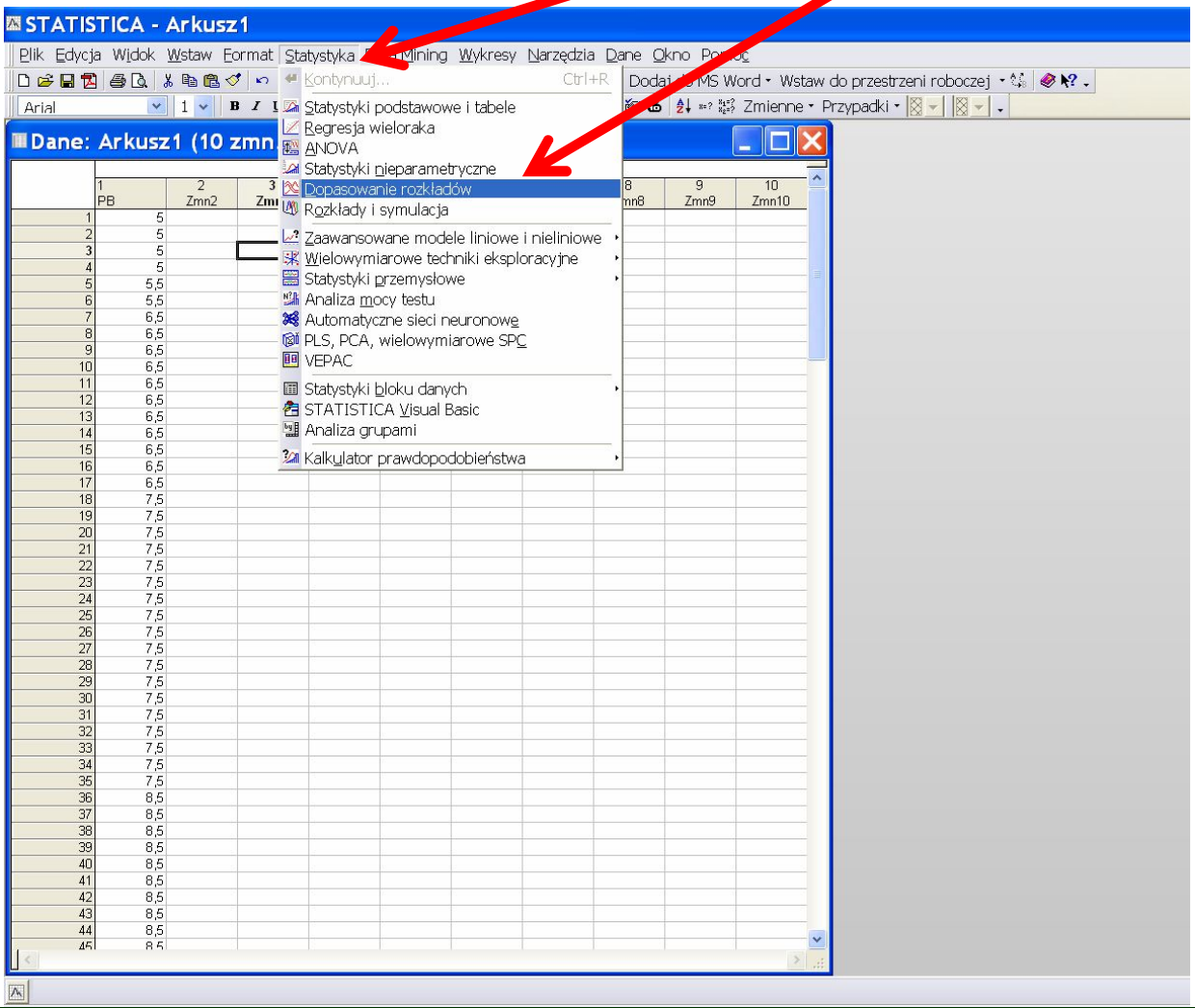
Jeżeli p-value > α, to hipotezę zerową odrzucamy, w przeciwnym przypadku hipotezy zerowej nie można odrzucić.

Obliczenia w pakiecie *STATISTICA*

W pliku **Parametr biochemiczny.xls** zaznaczamy zakres **A1:A193**, operacją **Kopiuj** umieszczamy je w schowku, uruchamiamy pakiet, klikamy w arkuszu danych w pierwszej komórce wybranej kolumny i z menu podręcznego wybieramy **Wklej z nagłówkami/Wklej z nazwami zmiennych**.

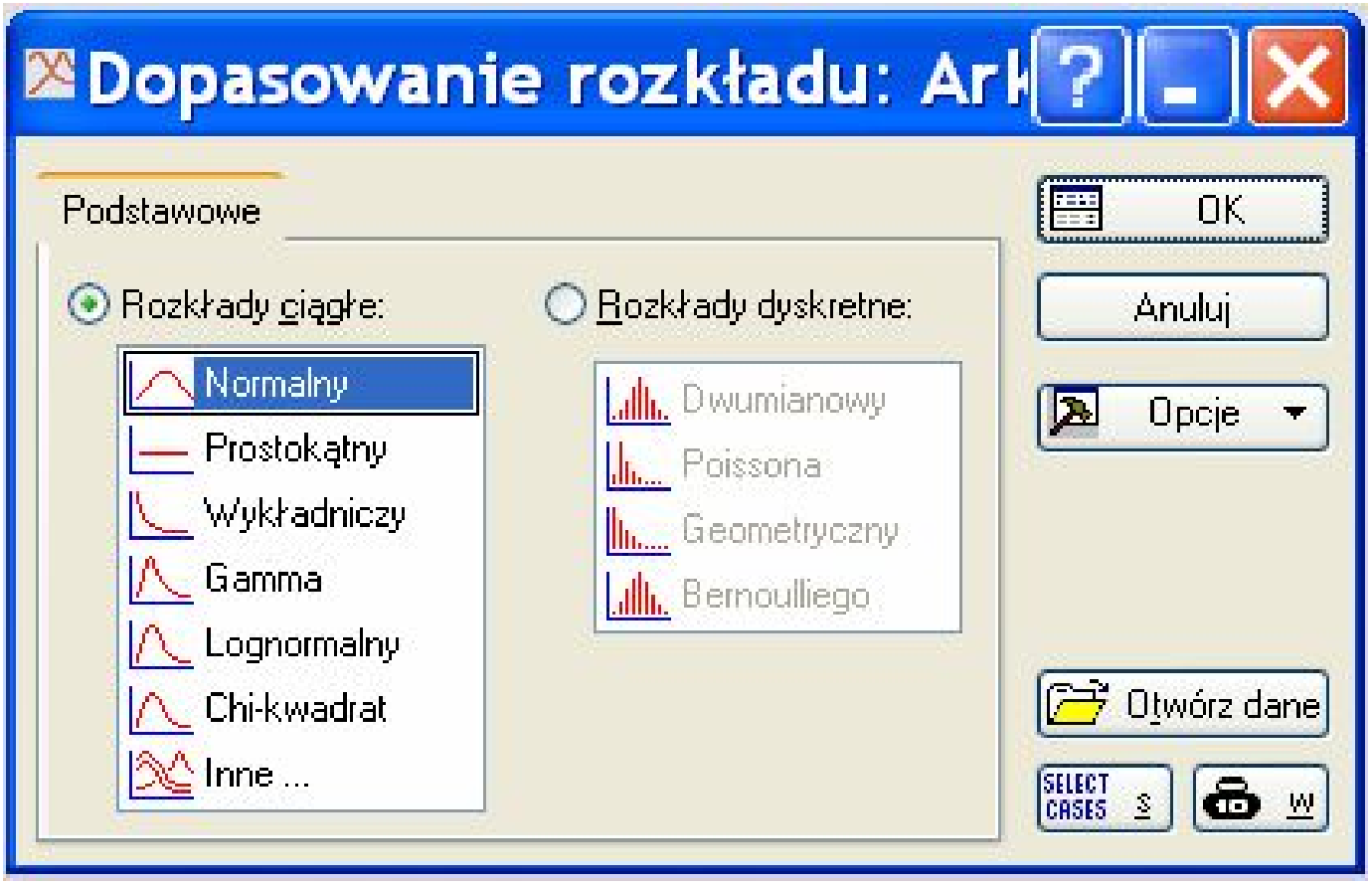
Obliczenia w pakiecie *STATISTICA*

Z menu głównego wybieramy opcję **Statystyka**, z podmenu **Dopasowanie rozkładów**.



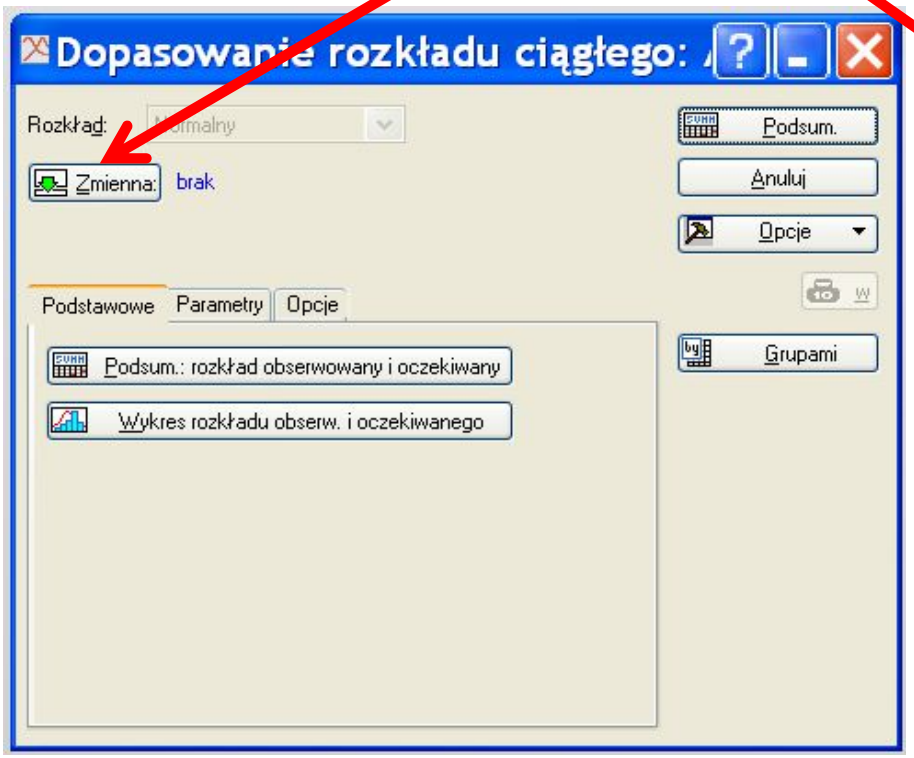
Obliczenia w pakiecie *STATISTICA*

W okienku **Dopasowanie rozkładu** z listy **Rozkłady ciągłe** wybieramy nazwę **Normalny**, **OK**.



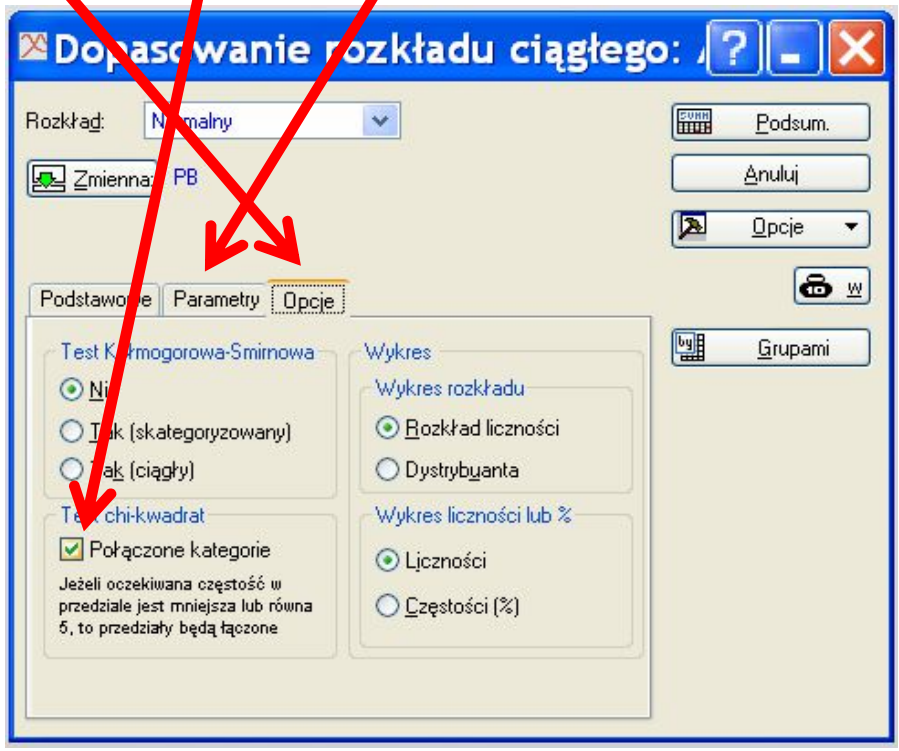
Obliczenia w pakiecie *STATISTICA*

W okienku **Dopasowanie rozkładu ciągłego** przyciskamy **Zmienne** i wprowadzamy nazwę kolumny z danymi **PB**, **OK**.



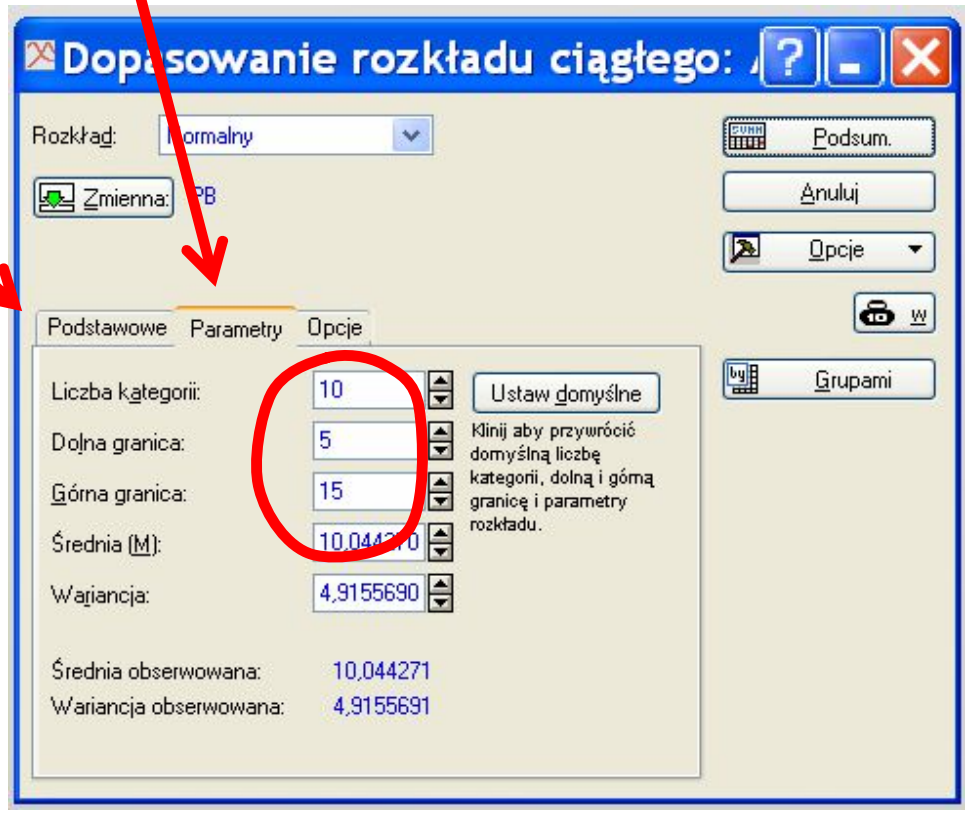
Obliczenia w pakiecie *STATISTICA*

W okienku **Dopasowanie rozkładu ciągłego** na karcie **Opcje** w polu **Test chi-kwadrat** zaznaczamy **Połączone kategorie**. Przechodzimy do karty **Parametry**.



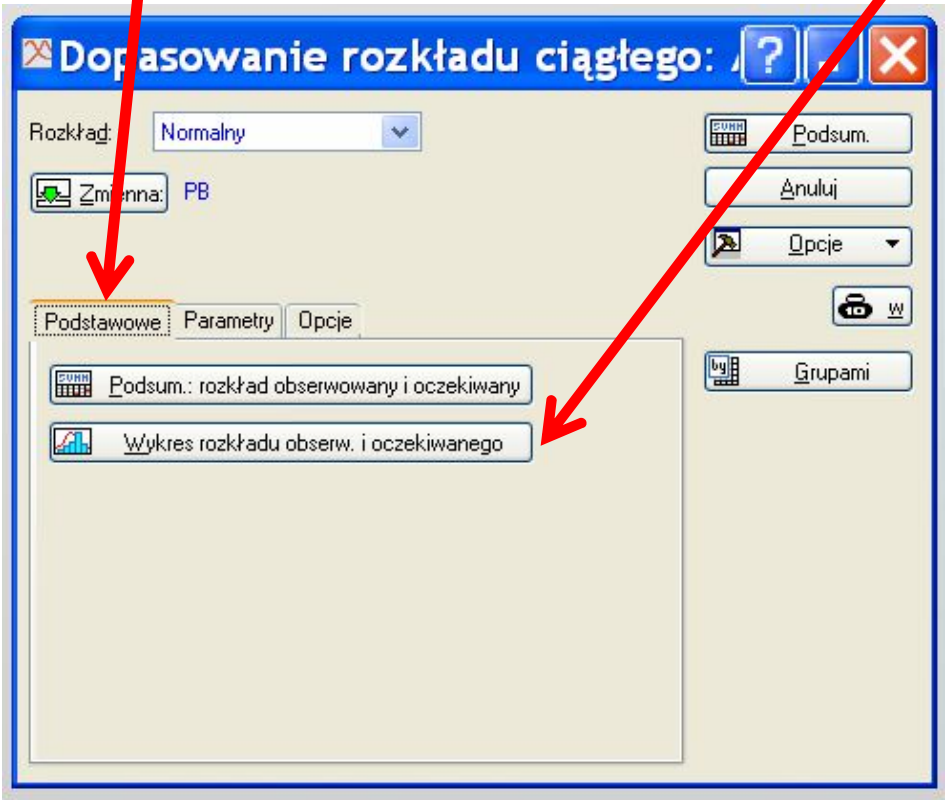
Obliczenia w pakiecie *STATISTICA*

Na karcie **Parametry** ustawiamy liczbę kategorii (czyli klas szeregu rozdzielczego) **10**, dolną granicę **5**, górną granicę **15** (średnią i wariancję pakiet liczy domyślnie). Przechodzimy do karty **Podstawowe**.



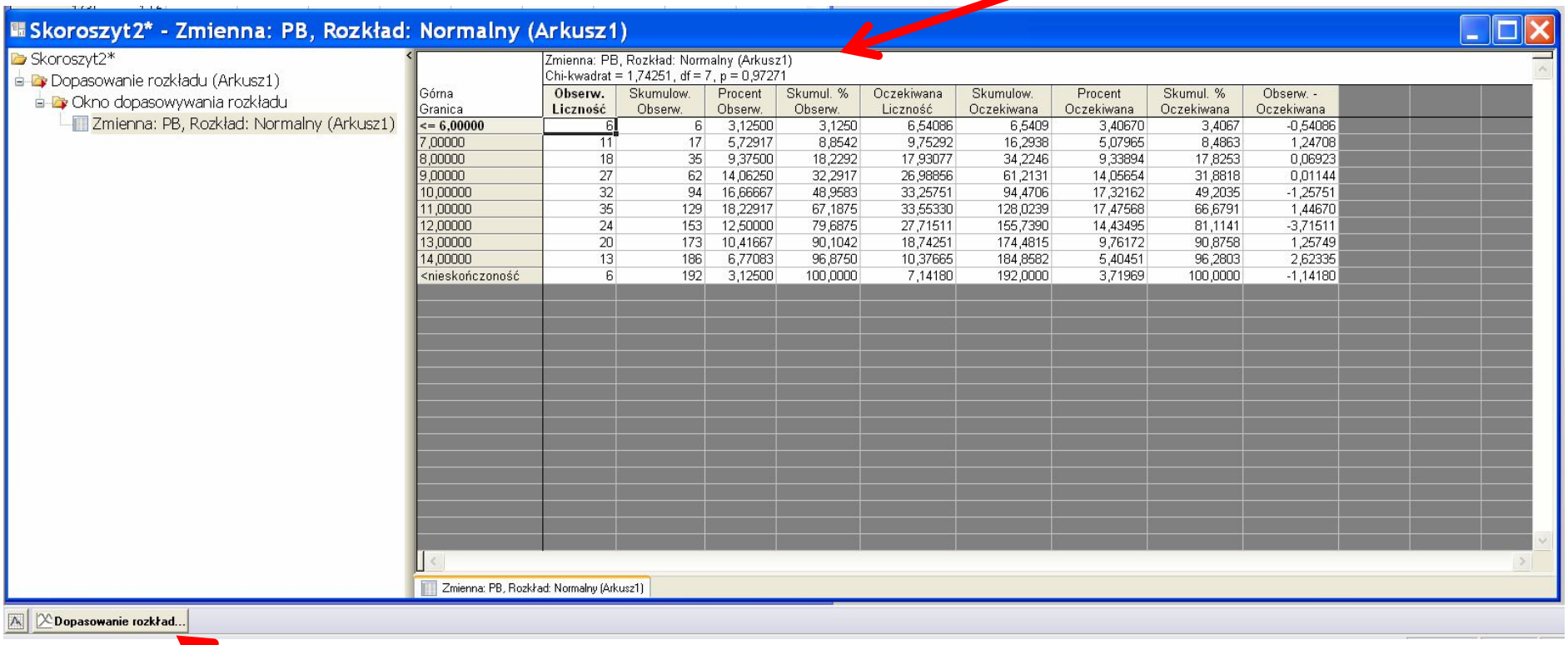
Obliczenia w pakiecie *STATISTICA*

Na karcie **Podstawowe** przyciskamy **Podsum.: rozkład obserwowany i oczekiwany** ….



Obliczenia w pakiecie *STATISTICA*

... i otrzymujemy szereg rozdzielczy, a w tytule jest wynik testu chi-kwadrat.

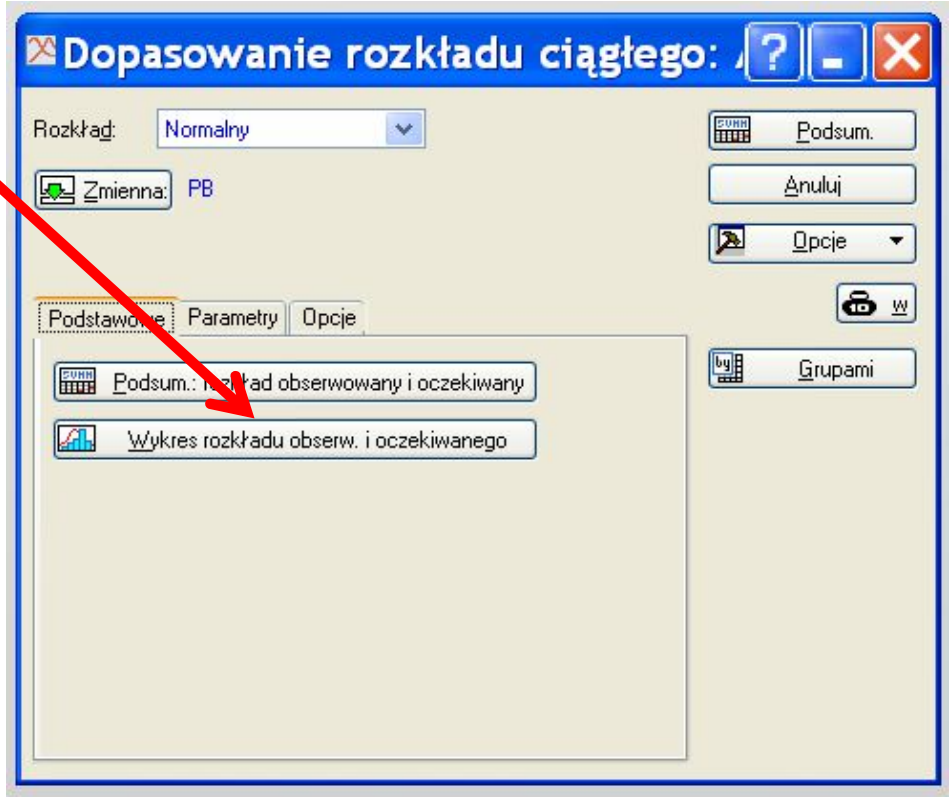


Górna Granica	Obserw. Liczność	Skumulow. Obserw.	Procent Obserw.	Skumul. % Obserw.	Oczekiwana Liczność	Skumulow. Oczekiwana	Procent Oczekiwana	Skumul. % Oczekiwana	Obserw. - Oczekiwana
<= 6,00000	6	6	3,12500	3,1250	6,54086	6,5409	3,40670	3,4067	-0,54086
7,00000	11	17	5,72917	8,8542	9,75292	16,2938	5,07965	8,4863	1,24708
8,00000	18	35	9,37500	18,2292	17,93077	34,2246	9,33894	17,8253	0,06923
9,00000	27	62	14,06250	32,2917	26,98856	61,2131	14,05654	31,8818	0,01144
10,00000	32	94	16,66667	48,9583	33,25751	94,4706	17,32162	49,2035	-1,25751
11,00000	35	129	18,22917	67,1875	33,55330	128,0239	17,47568	66,6791	1,44670
12,00000	24	153	12,50000	79,6875	27,71511	155,7390	14,43495	81,1141	-3,71511
13,00000	20	173	10,41667	90,1042	18,74251	174,4815	9,76172	90,8758	1,25749
14,00000	13	186	6,77083	96,8750	10,37665	184,8582	5,40451	96,2803	2,62335
<nieskończoność	6	192	3,12500	100,0000	7,14180	192,0000	3,71969	100,0000	-1,14180

Ponownie otwieramy okienko naszej analizy klikając na wizytówkę ...

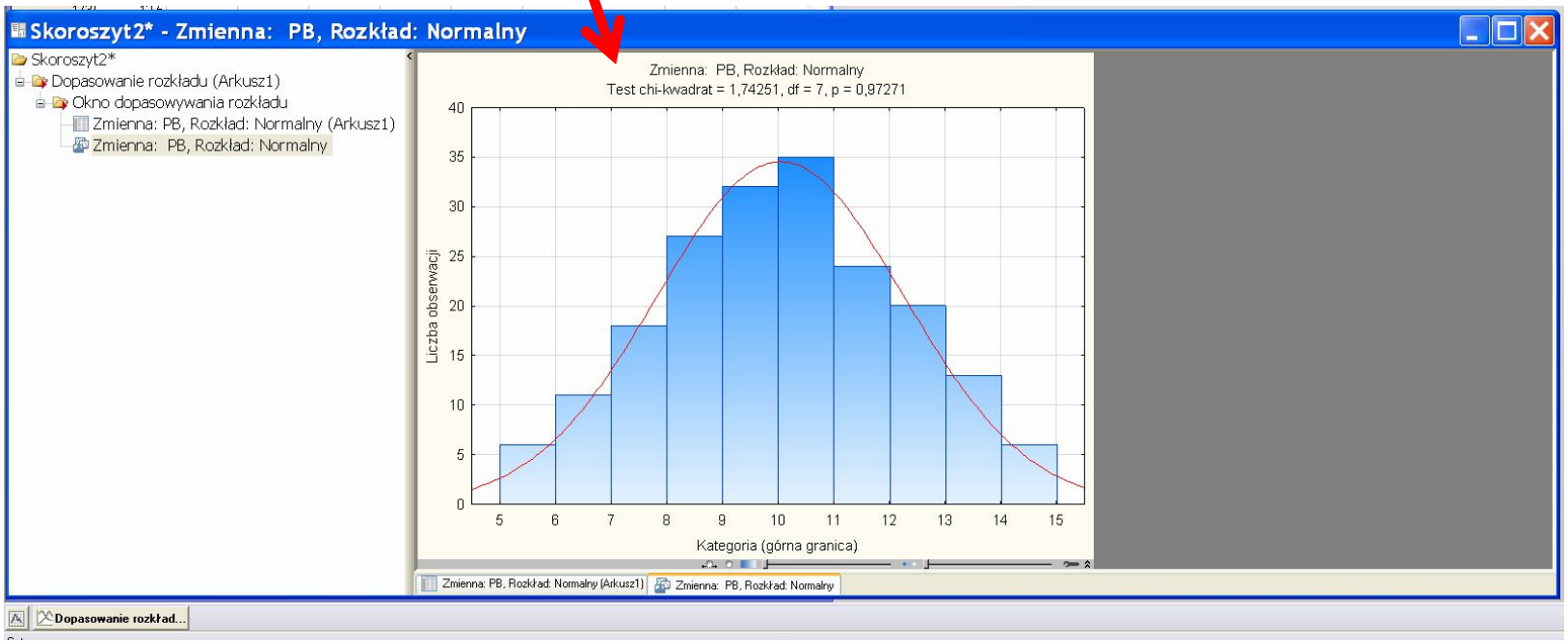
Obliczenia w pakiecie *STATISTICA*

... i na karcie **Podstawowe** przyciskamy **Wykres rozkładu obserw. i oczekiwanego** ....



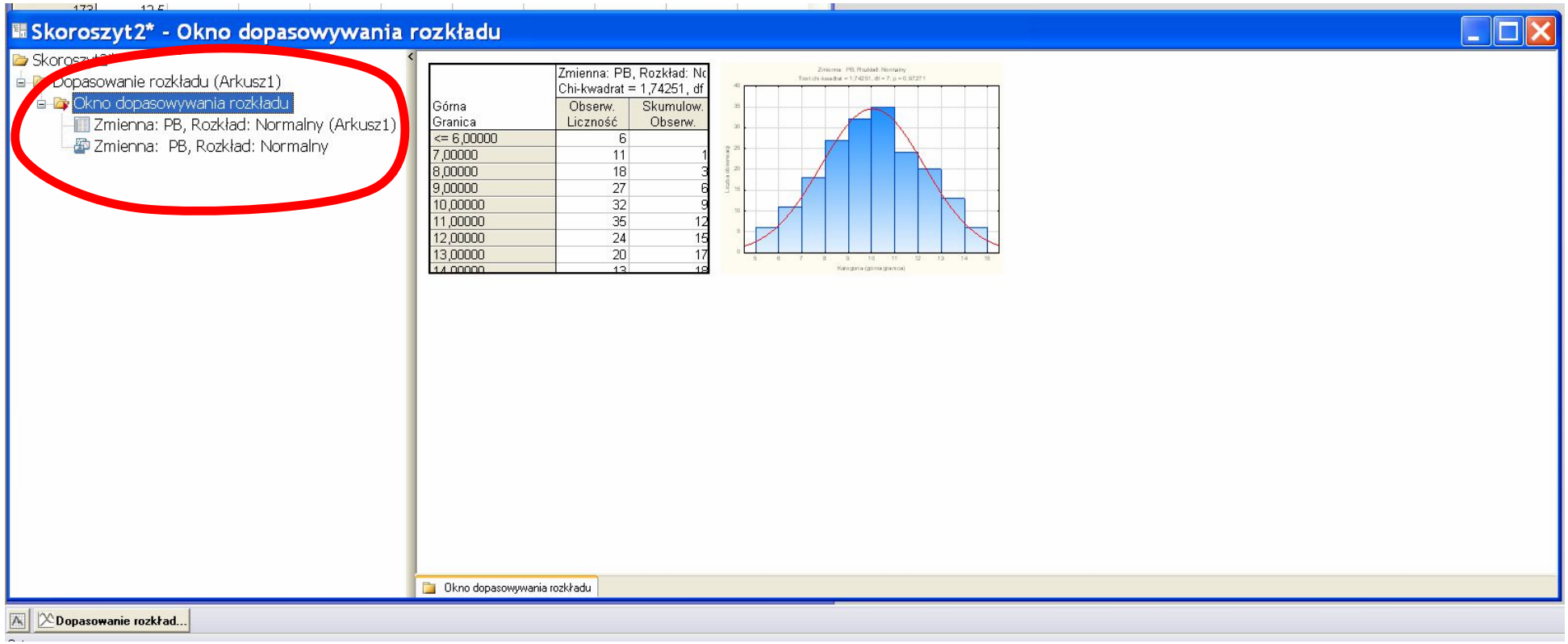
Obliczenia w pakiecie *STATISTICA*

... i otrzymujemy szereg rozdzielczy i krzywą normalną, a w tytule jest wynik testu chi-kwadrat.



Obliczenia w pakiecie *STATISTICA*

W oknie skoroszytu z wynikami są dostępne: wykres, tabela.



Badanie normalności rozkładu

Testy do badania normalności rozkładu w pakiecie *STATISTICA*:

- test chi-kwadrat zgodności
- test W Shapiro-Wilka
- test Kołmogorowa-Smirnowa (K-S test)
- test Lillieforsa

Test chi-kwadrat zgodności

- bada zgodność rozkładu empirycznego z wybranym rozkładem teoretycznym
- stosowany jest do danych w postaci szeregu rozdzielczego
- wymaga, aby liczebność próby przekraczała 50 elementów, a liczebności oczekiwane były większe niż 5 (jeśli ten warunek nie jest spełniony, należy łączyć klasy)

W pakiecie *STATISTICA*: menu Statystyka/Dopasowanie rozkładów.

Liczba klas w szeregu rozdzielczym

Liczba klas k zależy od liczebności próby n
i można ją przybliżeniu ustalić w oparciu o wzory:

$$k \approx \sqrt{n}$$

$$k \leq 5 \ln n$$

$$k \approx 1 + 3{,}322 \ln n$$

n	k
30-60	6 - 8
60-100	7 - 10
100-200	9 - 12
200-500	11 - 17
500-1500	16 – 25
przeważnie najwyżej 30	

Test W Shapiro-Wilka

- najbardziej polecany do badania normalności rozkładu (duża moc w porównaniu z innymi testami)
- zamiast porównania rozkładów p-stwa (jak chi-kwadrat), porównywuje wartości dystrybuanty
- w programie *STATISTICA* zastosowanao jego odmianę pozwalającą na testowanie bardzo dużych prób, jednak gdy próba przekracza 2000 elementów, test może dawać błędne wyniki; wówczas należy stosować test Lillieforsa lub chi-kwadrat

W pakiecie *STATISTICA*: menu Statystyka/Statystyki podstawowe i tabele/Statystyki opisowe/karta Normalność.

Test Kołmogorowa-Smirnowa i Lillieforsa

- zamiast porównania rozkładów p-stwa (jak chi-kwadrat), porównywuje wartości dystrybuanty
- wymaga znajomości średniej i odchylenia standardowego populacji, z której próba pochodzi (dlatego stosuje się go rzadko); gdy nie znamy wymaganych parametrów stosujemy test K-S z poprawką Lillieforsa

W pakiecie *STATISTICA*: menu Statystyka/Dopasowanie rozkładów/karta Opcje oraz menu Statystyka/Statystyki podstawowe i tabele/Statystyki opisowe/karta Normalność (K-S test).

Obliczenia w pakiecie *STATISTICA*

- Hipotezy zerowe dla wszystkich testów normalności mówią, że rozkład badany jest rozkładem normalnym (odrzucenie hipotezy zerowej oznacza, że badana cecha nie ma rozkładu normalnego).
- testów sprawdzających normalność nie dotyczy sytuacja, gdy zaznaczona jest opcja **Przedziały całkowitoliczbowe**.
- Przy tworzeniu histogramu zaznaczenie opcji **Przedziały całkowitoliczbowe** powoduje eliminacje wszystkich danych nie będących liczbami całkowitymi.

Estymacja przedziałowa średniej μ

Cecha X ma w populacji rozkład normalny,

$X \sim N(\mu, \sigma^2)$, μ, σ^2 – nieznane

Losujemy próbę: $x_1, x_2, \ldots, x_n$

Obliczamy parametry próby: $\bar{x}$, s^2, s

Wzór na przedział ufności dla μ

przy poziomie ufności $P=1-\alpha$

$$\mu \in \left(\bar{x} - t_{\alpha,v} \cdot \frac{s}{\sqrt{n}}; \bar{x} + t_{\alpha,v} \cdot \frac{s}{\sqrt{n}} \right)$$

$t_{\alpha, v}$ – wartość krytyczna z rozkładu t-Studenta

v – liczba stopni swobody, $v = n-1$

Estymacja przedziałowa średniej μ

Poziom ufności jest to p-stwo zdarzenia, że przedział ufności zawiera prawdziwą wartość średniej populacyjnej μ.

Powtórzenie. Przykład

Cecha X ma w populacji rozkład normalny, $X \sim N(\mu, \sigma^2)$, μ, σ^2 – nieznane. Wyznacz 95-procentowy przedział ufności dla średniej μ na podstawie próby:

191,2	193,0	195,1	184,3	197,6
200,8	194,2	198,7	189,5	200,2

Parametry próby:

$$n = 10 \quad \overline{x} = 194{,}46 \quad s^2 = 26{,}89 \quad s = 5{,}19$$

Wartość krytyczna z rozkładu t-Studenta:

$$t_{\alpha,\, v=n-1} = t_{0{,}05,\, 9} = 2{,}2622$$

95-proc. przedział ufności dla średniej μ: $(190{,}75\,;\,198{,}17)$

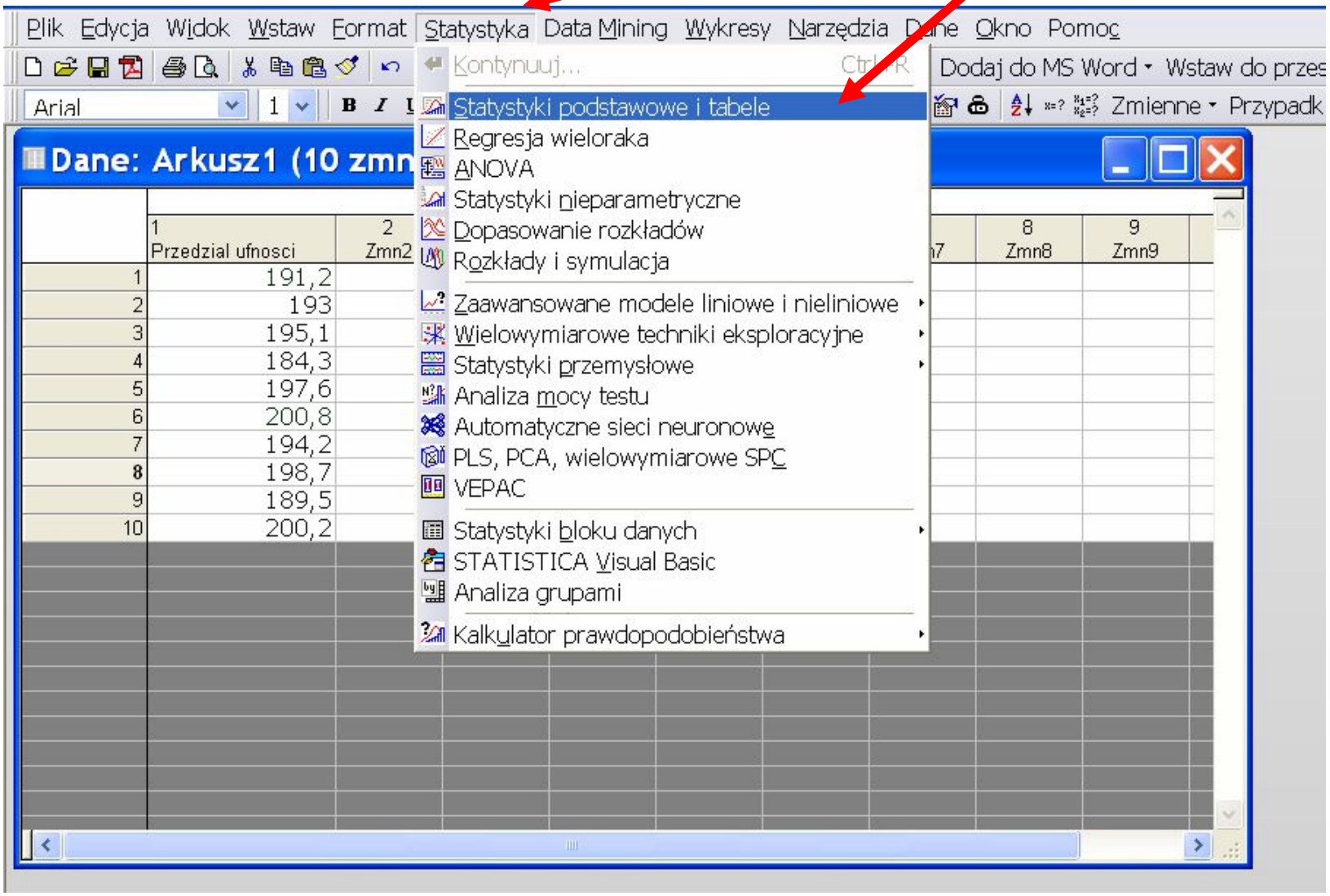
Obliczenia w pakiecie *STATISTICA*

W pakiecie STATISTICA przedział ufności dla średniej jest wyznaczany z tego samego wzoru.

1. Dane **Przedzial ufnosci** z arkusza wklejamy do schowka poleceniem **Kopiuj**.
2. Uruchamiamy pakiet, zamykamy okienko powitalne, klikamy w pierwszej komórce dowolnej kolumny arkusza danych, klikamy prawym przyciskiem myszy na tle nagłówka kolumny i z menu podręcznego wybieramy **Wklej z nagłówkami /Wklej z nazwami zmiennych.**

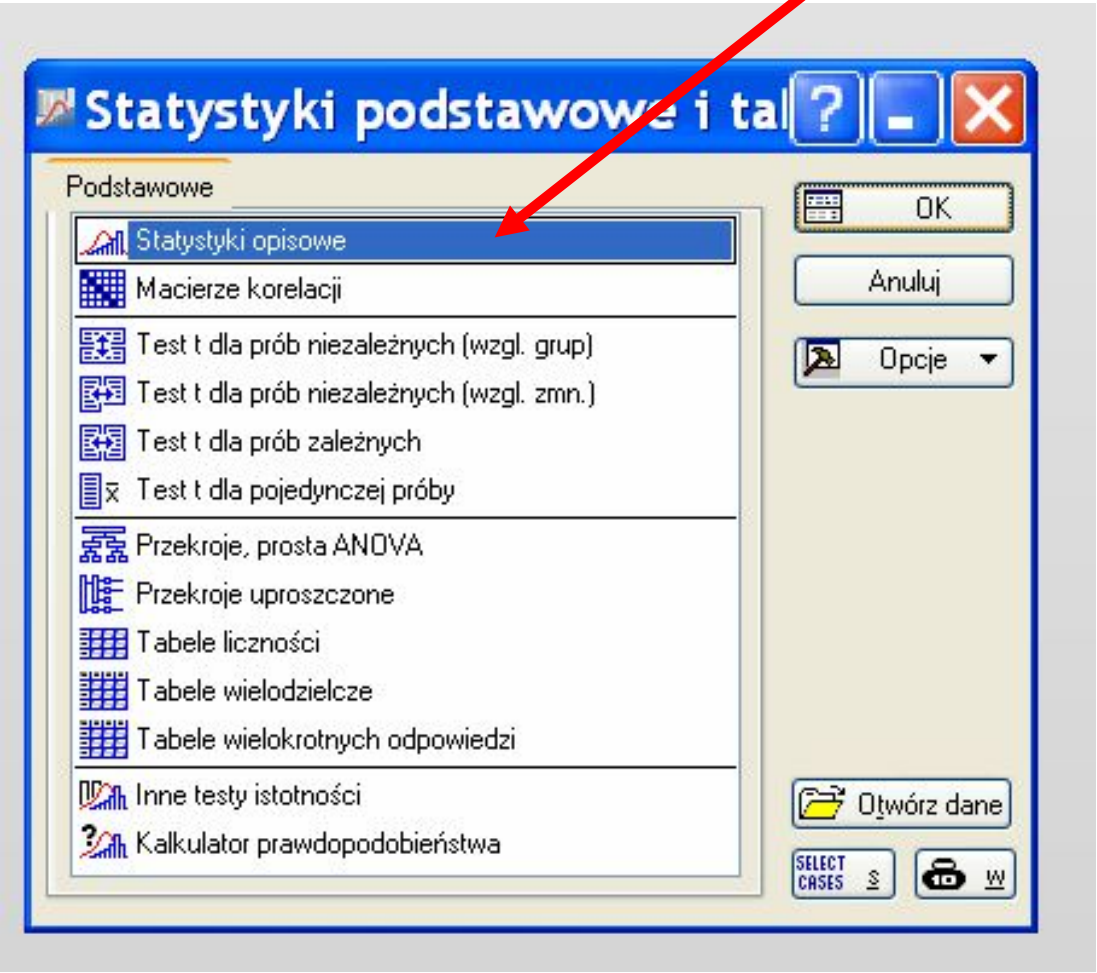
Obliczenia w pakiecie *STATISTICA*

3. Z menu głównego wybieramy **Statystyka**, z podmenu **Statystyki podstawowe i tabele.**



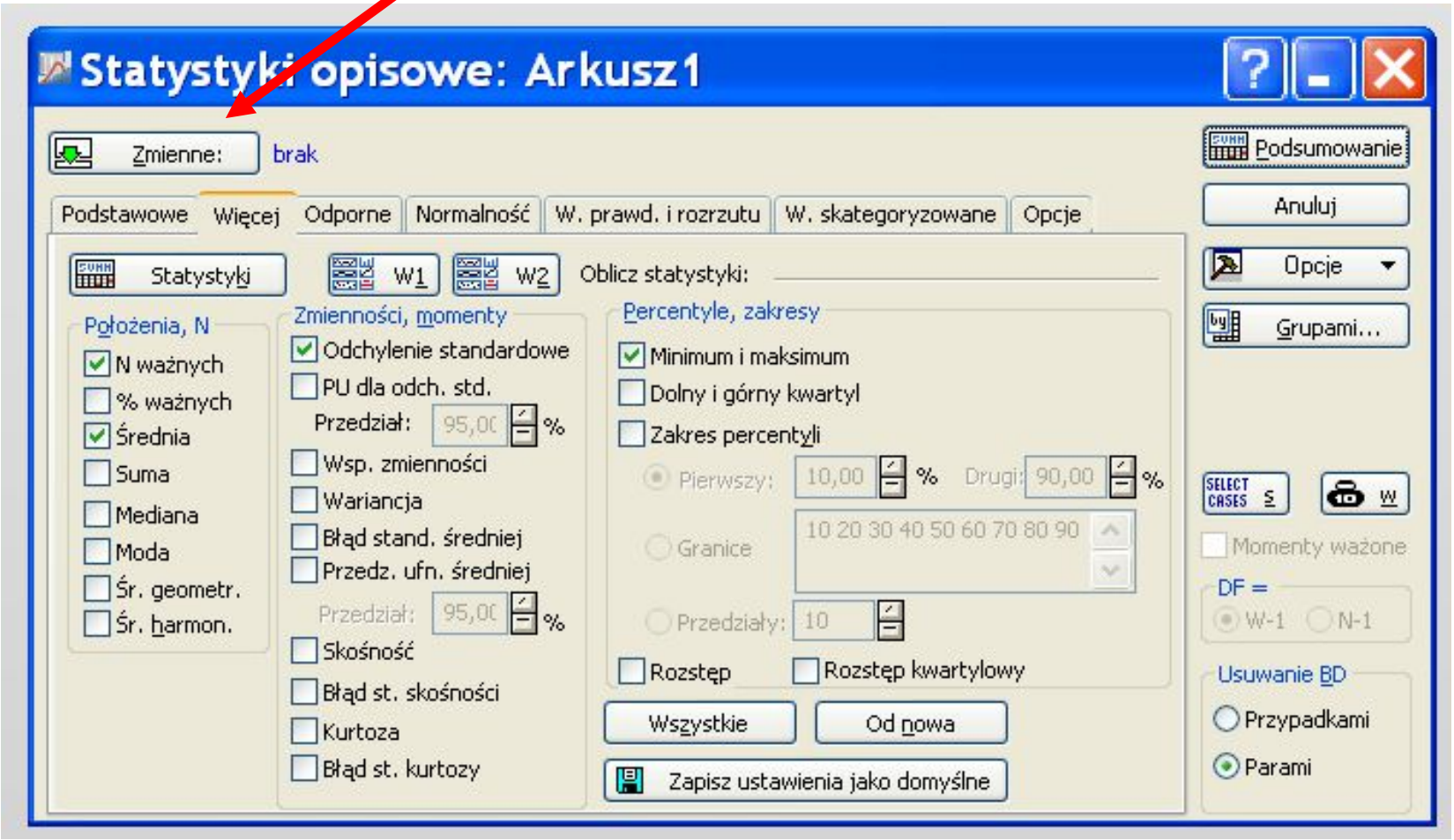
Obliczenia w pakiecie *STATISTICA*

4. Z kolejnego podmenu **Statystyki opisowe**, **OK.**



Obliczenia w pakiecie *STATISTICA*

5. Przyciskamy **Zmienne, ….**

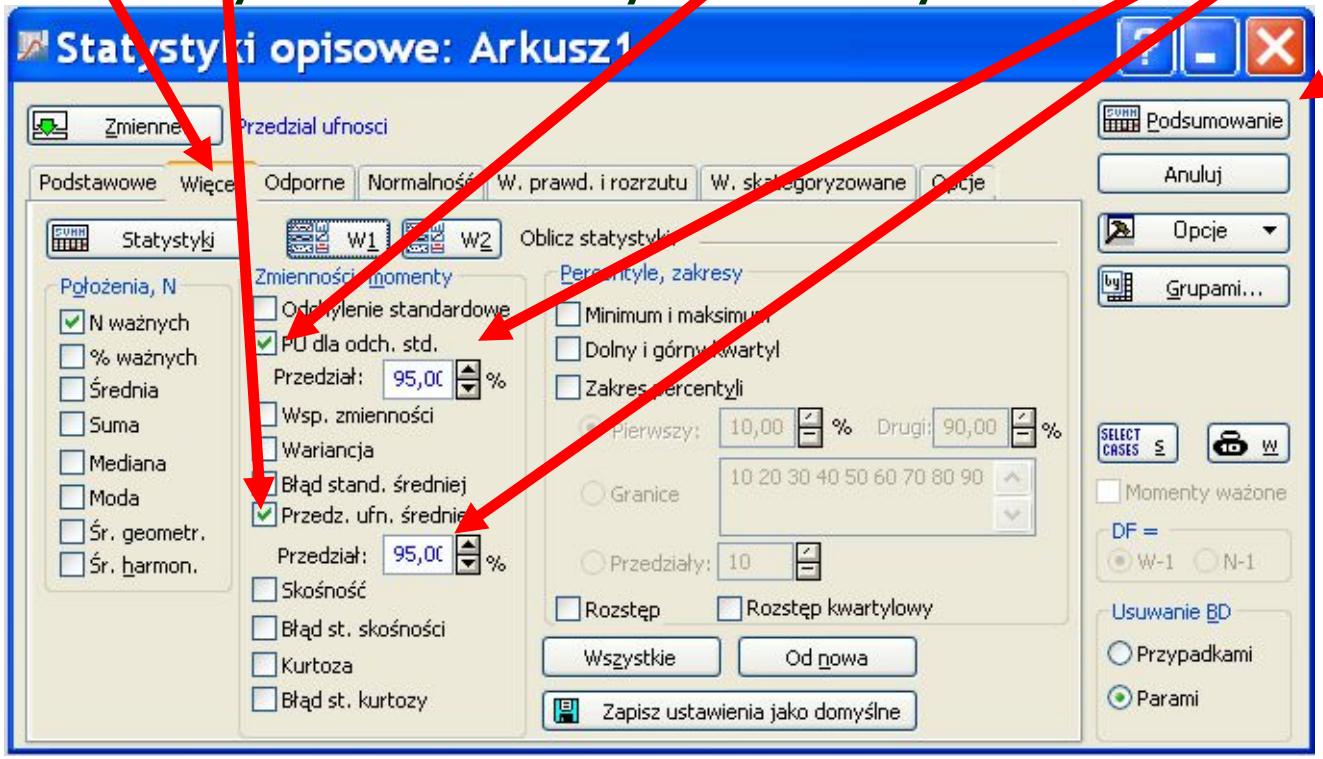


Obliczenia w pakiecie *STATISTICA*

6. …. wybieramy z listy kolumnę z danymi **Przedzial ufnosci**, **OK**.

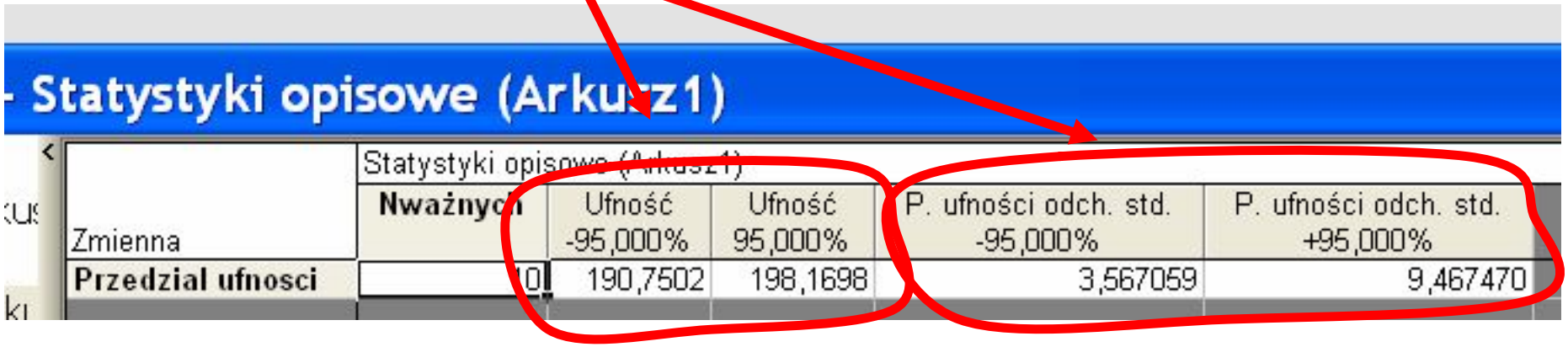
Obliczenia w pakiecie *STATISTICA*

7. W oknie **Statystyki opisowe** wybieramy kartę **Więcej** i zaznaczamy na niej pola: **PU dla odch. std.**, **Przedz.ufn. średniej**. Pod każdym z nich można wpisać poziom ufności (domyślnie jest 95%). Wpisujemy tylko liczbę **95**, symbol % pokazany jest obok na szarym tle. Przyciskamy **Podsumowanie**.



Obliczenia w pakiecie *STATISTICA*

8. Pojawia się arkusz z wynikami. Trzeba zgadnąć, że tu są krańce 95-proc. przedziału ufności dla średniej populacyjnej (opis nie jest jasny), a tu krańce przedziału ufności dla odchylenia standardowego.



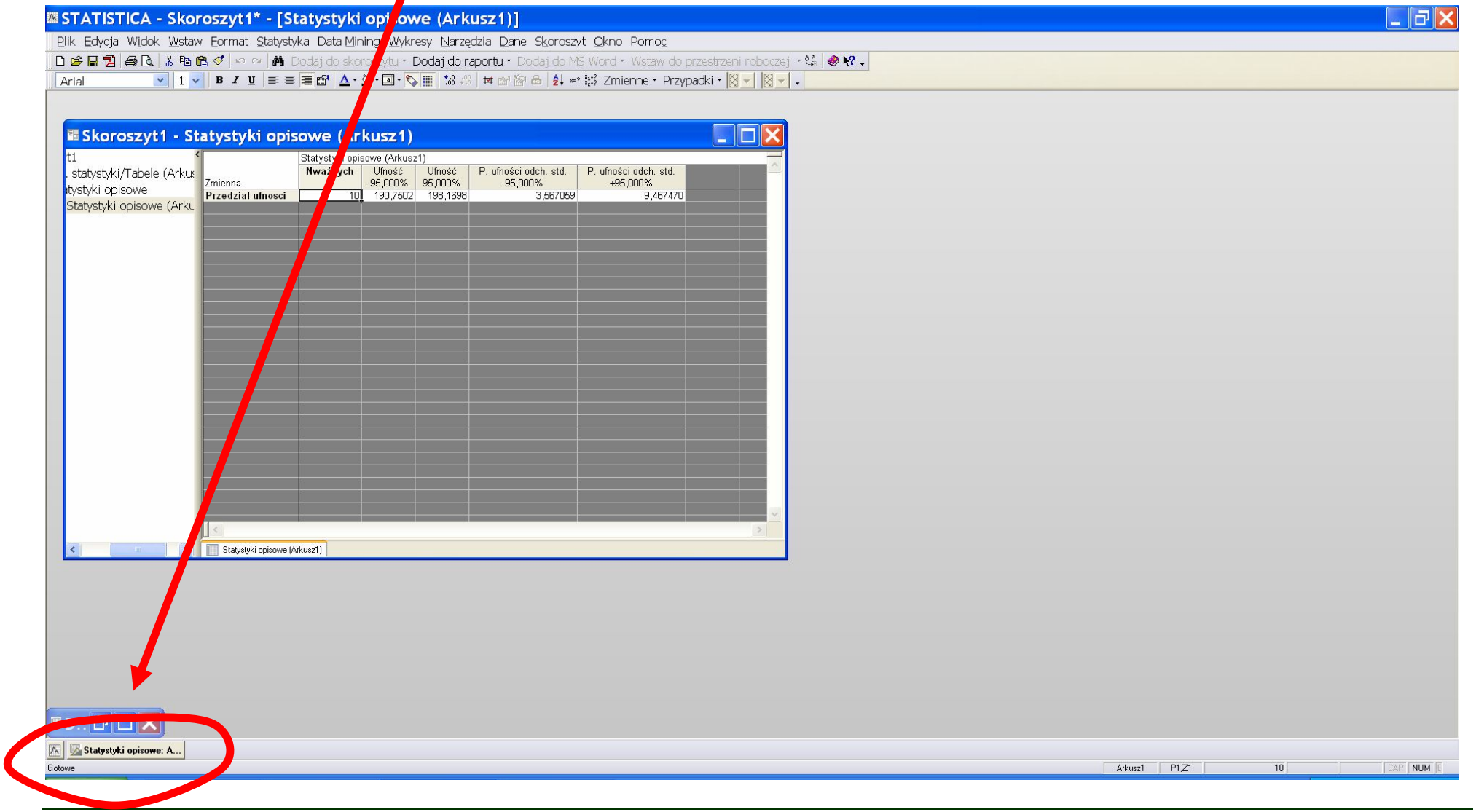
Statystyki opisowe (Arkusz1)

Zmienna	Statystyki opisowe (Arkusz1)				
	Nważnych	Ufność -95,000%	Ufność 95,000%	P. ufności odch. std. -95,000%	P. ufności odch. std. +95,000%
Przedział ufności	10	190,7502	198,1698	3,567059	9,467470

Wykresy z przedziałmi ufności można wykonać wybierając z menu **Wykresy**, podmenu **Wykresy średnia i błędy**.

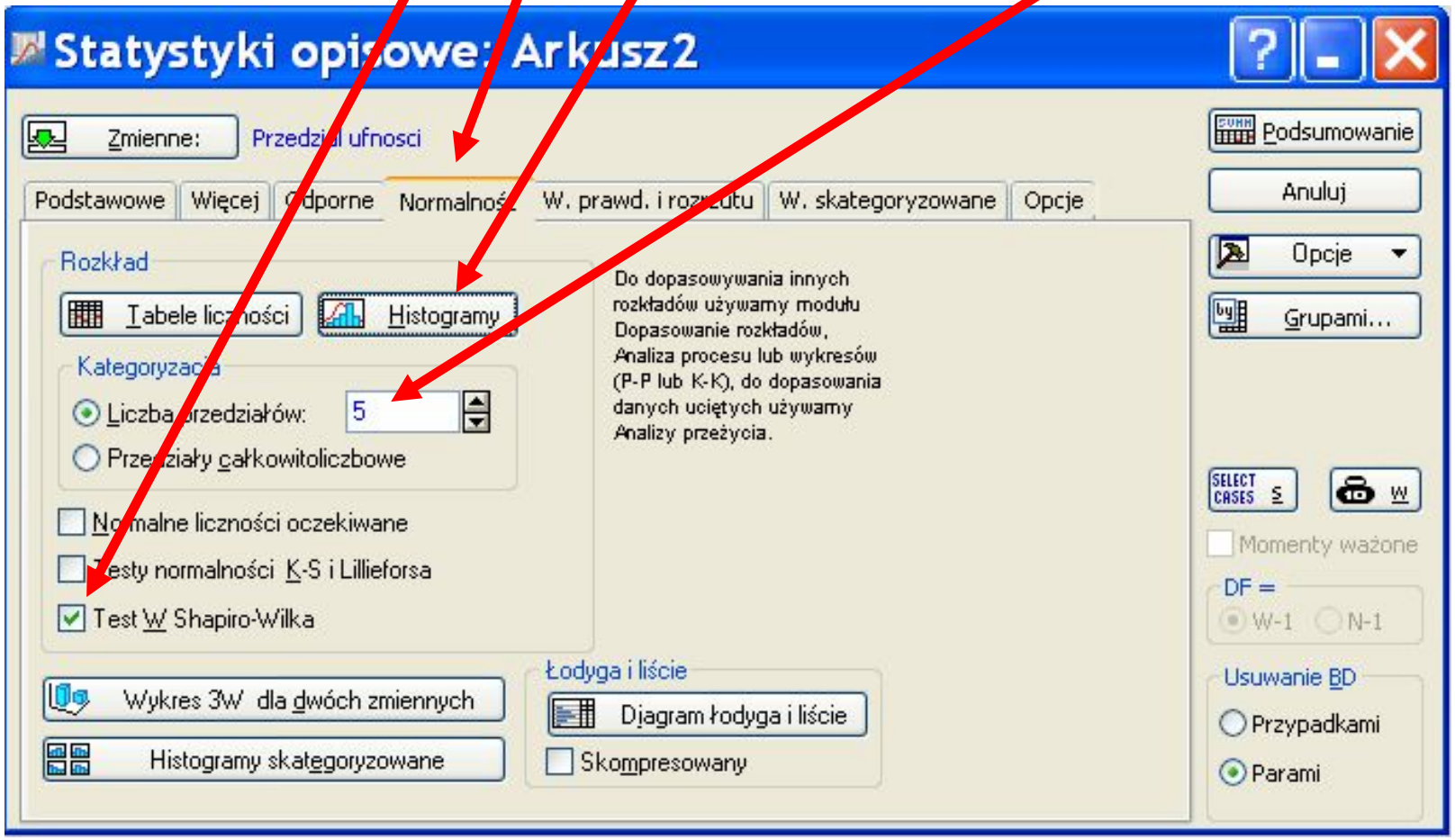
Obliczenia w pakiecie *STATISTICA*

9. Okno **Statystyki opisowe** zostało zminimalizowane. Klikamy na nim.



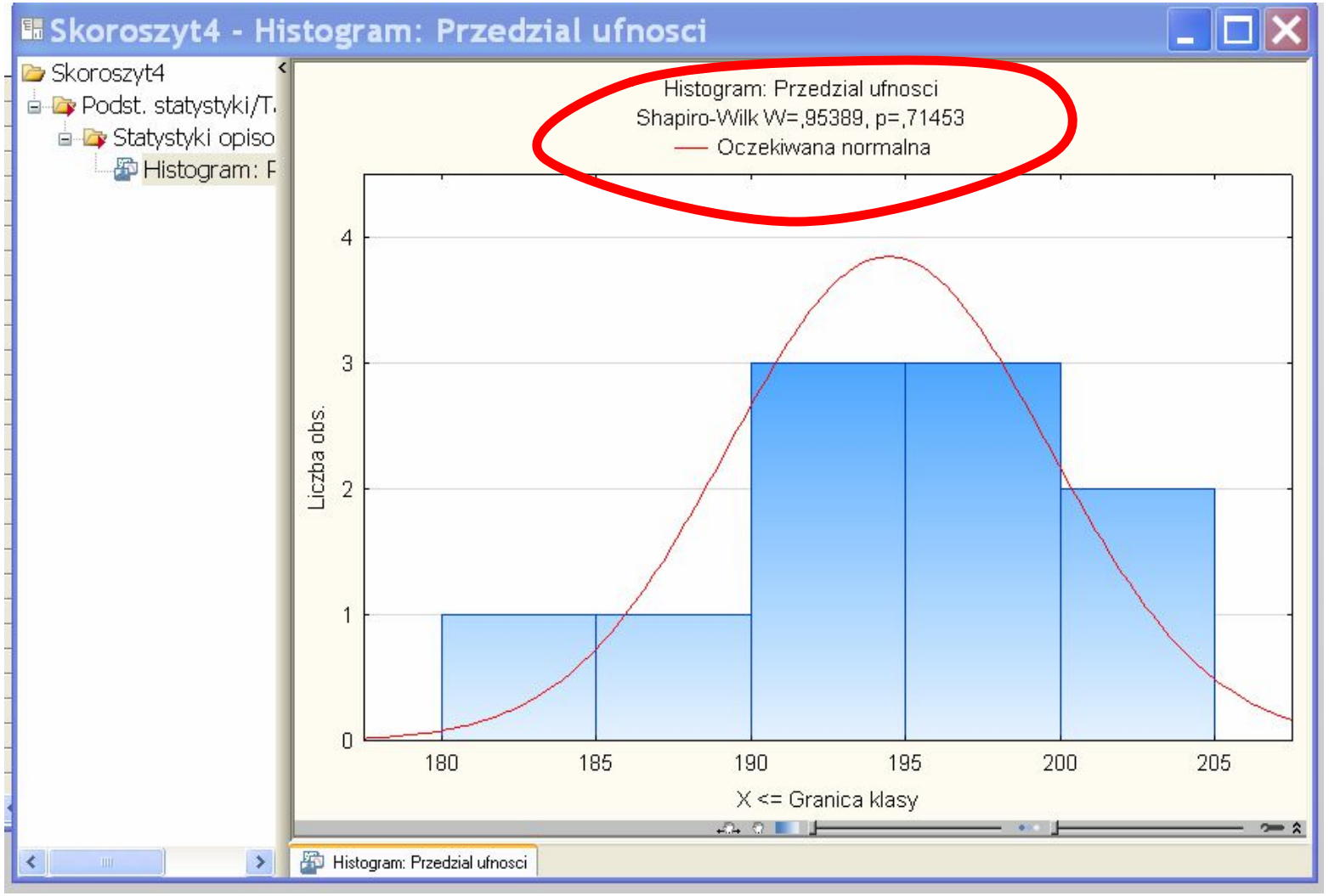
Obliczenia w pakiecie *STATISTICA*

10. Na karcie **Normalność** zaznaczamy: **5** przedziałów, **test W Shapiro-Wilka**, przyciskamy **Histogramy**



Obliczenia w pakiecie *STATISTICA*

11. Nad wykresem wyświetlony jest wynik testowania.



Test *t* dla dwóch prób niezależnych

Przykład. U zdrowych i chorych na chorobę niedokrwienną serca zbadano całkowite stężenie cholesterolu. Czy w grupie chorych średnie stężenie cholesterolu jest różne od średniego stężenia w grupie kontrolnej (zdrowych)?

Dane w pliku cholesterol.xls

Test *t* dla dwóch prób niezależnych

X_1 – stężenie cholesterolu u zdrowych

X_2 – stężenie cholesterolu u chorych

Założenia

Cecha $X_1 \sim N(\mu_1, \sigma^2)$, cecha $X_2 \sim N(\mu_2, \sigma^2)$,

μ_1, μ_2, σ^2 - nieznane

Zbadano $n_1 = 37$ elementową próbę zdrowych ludzi oraz $n_2 = 33$ elementową próbę chorych.

Próba 1: $\bar{x}_1 = 5{,}80, \quad s_1 = 0{,}38, \quad n_1 = 37$

Próba 1: $\bar{x}_2 = 6{,}42, \quad s_2 = 0{,}47, \quad n_2 = 33$

Test *t* dla dwóch prób niezależnych

Hipoteza zerowa $H_0: \mu_1 = \mu_2$

Hipoteza alternatywna $H_1: \mu_1 \neq \mu_2$

Test t-Studenta, poziom istotności α

Funkcja testowa:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_r}$$

gdzie:

$$s_r = \sqrt{s_e^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$ błąd stand. różnicy średnich

$$s_e^2 = \frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{n_1 + n_2 - 2}$$ wspólna wariancja

Test *t* dla dwóch prób niezależnych cd.

Wnioskowanie

Jeżeli $|t| > t_{\alpha,\nu}$ to hipotezę H_0 odrzucamy,

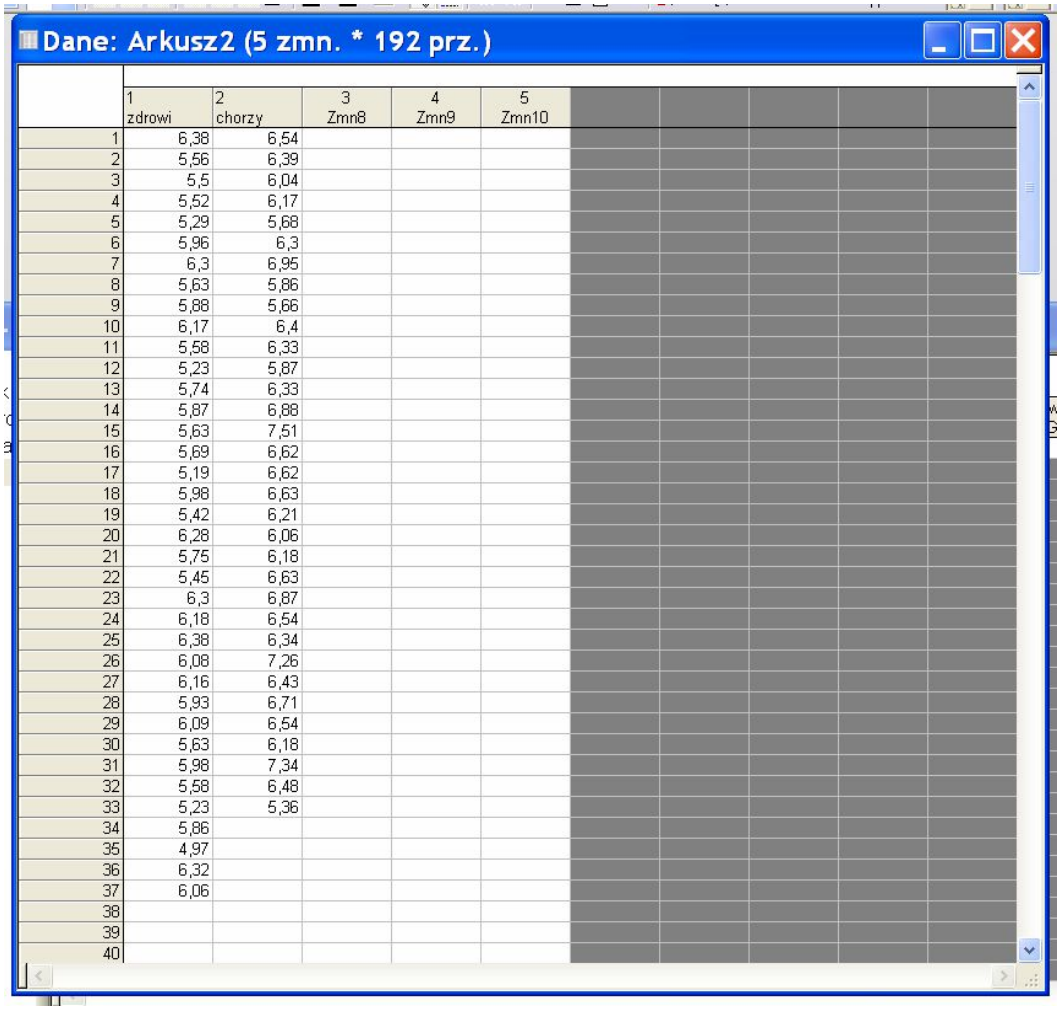
w przeciwnym przypadku H_0 nie można odrzucić.

$$\nu = n_1 + n_2 - 2$$

ν – liczba stopni swobody

Obliczenia w pakiecie *STATISTICA*

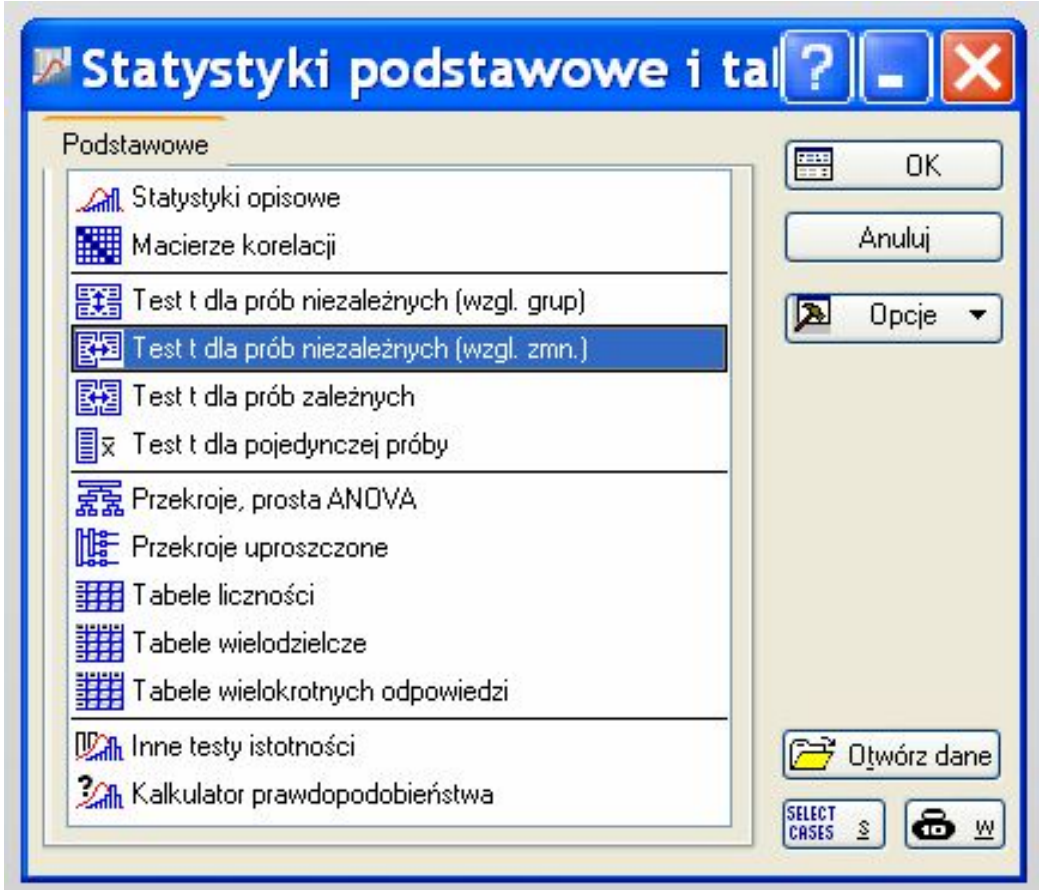
Kopiujemy dane z arkusza **cholesterol.xls**, wklejamy do arkusza *STATISTIKI*



	1 zdrowi	2 chorzy	3 Zmn8	4 Zmn9	5 Zmn10
1	6,38	6,54			
2	5,56	6,39			
3	5,5	6,04			
4	5,52	6,17			
5	5,29	5,68			
6	5,96	6,3			
7	6,3	6,95			
8	5,63	5,86			
9	5,88	5,66			
10	6,17	6,4			
11	5,58	6,33			
12	5,23	5,87			
13	5,74	6,33			
14	5,87	6,88			
15	5,63	7,51			
16	5,69	6,62			
17	5,19	6,62			
18	5,98	6,63			
19	5,42	6,21			
20	6,28	6,06			
21	5,75	6,18			
22	5,45	6,63			
23	6,3	6,87			
24	6,18	6,54			
25	6,38	6,34			
26	6,08	7,26			
27	6,16	6,43			
28	5,93	6,71			
29	6,09	6,54			
30	5,63	6,18			
31	5,98	7,34			
32	5,58	6,48			
33	5,23	5,36			
34	5,86				
35	4,97				
36	6,32				
37	6,06				
38					
39					
40					

Obliczenia w pakiecie *STATISTICA*

Wybieramy z menu **Statystyka/Statystyki podstawowe i tabele**, opcja **Test t dla prób niezależnych (wzgl. zmn.)**, **OK**.



Obliczenia w pakiecie *STATISTICA*

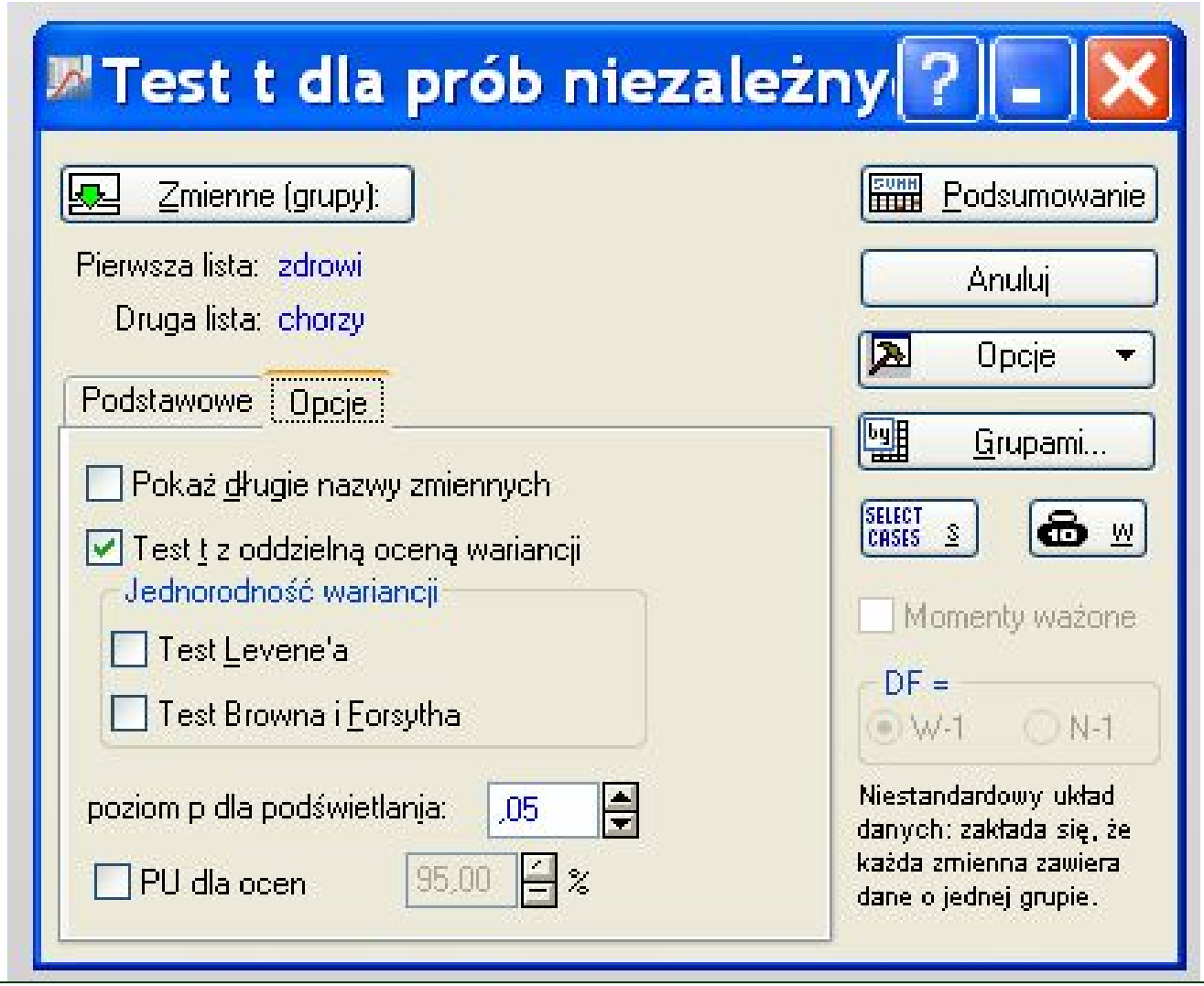
Klikamy przycisk **Zmienne (grupy)**.

Obliczenia w pakiecie *STATISTICA*

Z jednej listy wybieramy **zdrowi**, z drugiej **chorzy**, **OK**.

Obliczenia w pakiecie *STATISTICA*

Na karcie **Opcje** wybieramy **test t z oddzielną estymacja wariancji**.

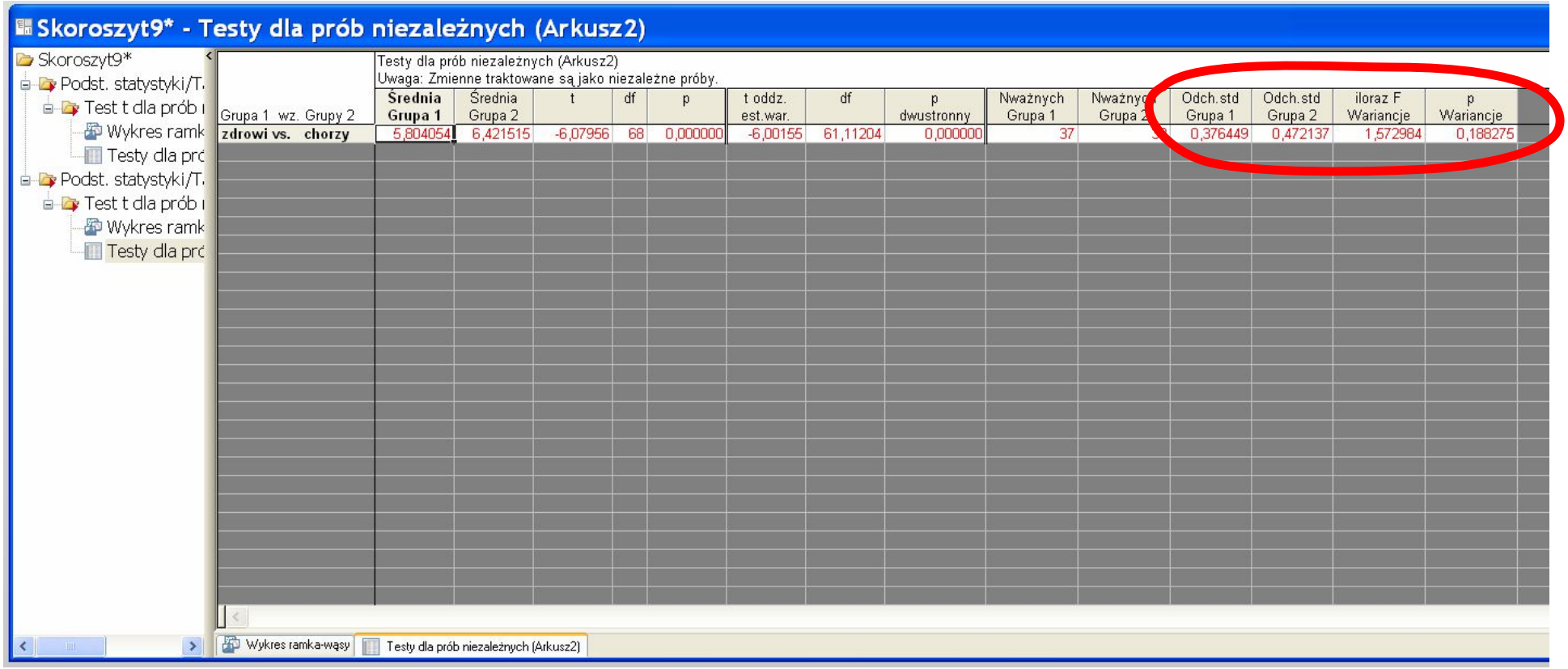


Obliczenia w pakiecie *STATISTICA*

Na karcie **Podstawowe** wybieramy **Podsumowanie: testy t**.

Obliczenia w pakiecie *STATISTICA*

Arkusz wyników czytamy od końca



Obliczenia w pakiecie *STATISTICA*

Tu zapisano wynik porównania dwóch wariancji!

	Nważnych Grupa 1	Nważnych Grupa 2	Odch.std Grupa 1	Odch.std Grupa 2	iloraz F Wariancje	p Wariancje
)	37	33	0,376449	0,472137	1,572984	0,188275

Porównanie dwóch wariancji

X_1 – stężenie cholesterolu u zdrowych

X_2 – stężenie cholesterolu u chorych

Zbadano n_1 =37 elementową próbę zdrowych ludzi oraz n_2 =33 elementową próbę chorych.

Założenia

Cecha $X_1 \sim N(\mu_1, \sigma_1^2)$, cecha $X_2 \sim N(\mu_2, \sigma_2^2)$,

μ_1, μ_2, σ_1^2, σ_2^2 - nieznane

Hipoteza zerowa $H_0: \sigma_1^2 = \sigma_2^2$

Hipoteza alternatywna $H_1: \sigma_1^2 \neq \sigma_2^2$

Test F-Fishera, poziom istotności α

Porównanie dwóch wariancji

Funkcja testowa:

$$F = \frac{\max(s_1^2, s_2^2)}{\min(s_1^2, s_2^2)}$$

<u>Wnioskowanie</u>

Jeżeli $F > F_{\alpha/2,\, \upsilon \text{ licz},\, \upsilon \text{ mian}}$, to hipotezę H_0 odrzucamy, w przeciwnym przypadku H_0 nie można odrzucić.

υ licz – liczba stopni swobody dla licznika, υ mian - liczba stopni swobody dla mianownika, $\upsilon_i = n_i - 1$

Obliczenia w pakiecie *STATISTICA*

Tu zapisano wynik porównania dwóch wariancji!

	Nważnych Grupa 1	Nważnych Grupa 2	Odch.std Grupa 1	Odch.std Grupa 2	iloraz F Wariancje	p Wariancje
)	37	33	0,376449	0,472137	1,572984	0,188275

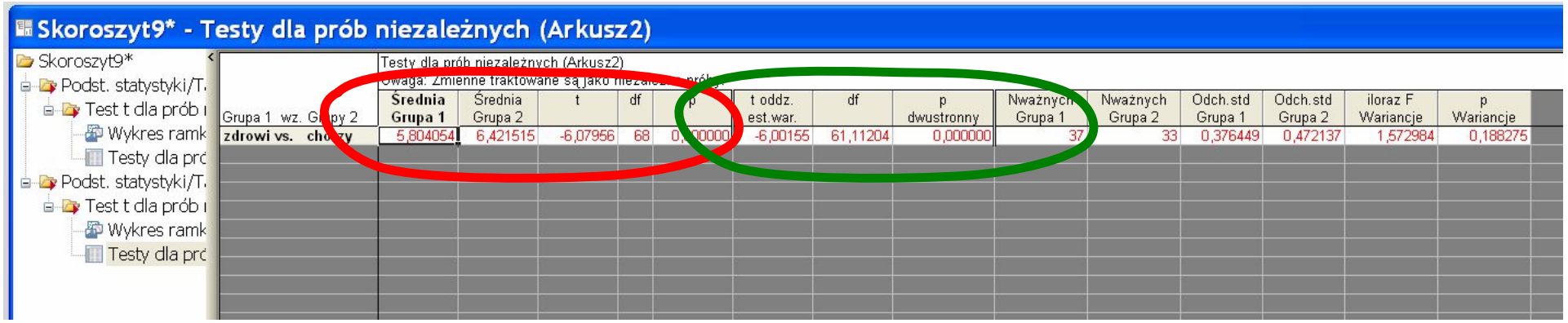
Wartość funkcji testowej F=1,57, *p*-value=0,19. Wariancje w grupie chorych i zdrowych nie różnią się istotnie.

Wnioskowanie równoważne

Jeżeli wartość-$p < \alpha$, to hipotezę H_0 odrzucamy, w przeciwnym przypadku H_0 nie można odrzucić.

Obliczenia w pakiecie *STATISTICA*

Wariancje są równe, więc wynik testu t bierzemy z miejsca zaznaczonego na czerwono. Gdyby wariancje były różne, wynik testu t należałoby czytać z miejsca zaznaczonego na zielono (test Cochrana-Coxa).



Wartość funkcji testowej t=-6,08, *p*-value<0,000000. Średnie w grupie chorych i zdrowych różnią się istotnie.

Test U Manna-Whitneya

Jest nieparametrycznym odpowiednikiem testu t-Studenta dla grup niezależnych.

Zastosowania

- gdy dane są mierzalne, ale nie mają rozkładu normalnego
- gdy dane nie są mierzalne, tylko typu porządkowego; nie można obliczyć średniej, a miarą położenia jest mediana

Test U Manna-Whitneya

Przykład. W dwóch grupach chorych na pewną chorobę neurologiczną przeprowadzono badania stężenia adrenaliny w surowicy krwi. Zebrane wyniki przedstawia tabela. Czy można przyjąć, że stężenie adrenaliny w obu grupach jest jednakowe?

Dane w tabeli i pliku adrenalina.xls.

Badana cecha w grupie 1 nie ma rozkładu normalnego – do sprawdzenia testem Shapiro-Wilka.

Nie można zastosować testu t-Studenta.

Test U Manna-Whitneya

Mamy dwie próby: n_1 elementową i n_2 elementową pobrane z populacji, w których badana cecha ma rozkłady typu ciągłego i dane można rozpatrywać w skali porządkowej.

Hipoteza zerowa

próby pochodzą z tej samej populacji (z populacji o równych medianach)

Hipoteza alternatywna

próby nie pochodzą z tej samej populacji

Test U-Manna-Whitneya, poziom istotności α

Test U Manna-Whitneya

Funkcja testowa:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

gdzie R_1 oznacza sumę rang przyznanych wartościom pierwszej próby.
Ta funkcja testowa ma rozkład podawany w tablicach statystycznych

Test U Manna-Whitneya

Każda liczba z połączonych grup dostanie rangę (numer) według kolejności rosnącej.

Grupa I	Stężenie	Ranga	Grupa II	Stężenie	Ranga
1	14,34		2	5,33	**1**
1	20,33		2	22,50	
1	18,79		2	11,74	
1	8,22	**4**	2	7,39	**2**
1	31,50		2	12,34	
1	12,08		2	13,22	
1	22,00		2	8,53	**5**
1	9,22		2	22,80	
1	19,50		2	12,70	
1	78,89		2	7,78	**3**
1	30,48		2	9,63	
1	45,86		2	8,90	

Test U Manna-Whitneya

Rangi zostały przydzielone.

Grupa I	Stężenie	Ranga	Grupa II	Stężenie	Ranga
1	14,34	**14**	2	5,33	**1**
1	20,33	**17**	2	22,50	**19**
1	18,79	**15**	2	11,74	**9**
1	8,22	**4**	2	7,39	**2**
1	31,50	**22**	2	12,34	**12**
1	12,08	**10**	2	13,22	**13**
1	22,00	**18**	2	8,53	**5**
1	9,22	**7**	2	22,80	**20**
1	19,50	**16**	2	12,70	**11**
1	78,89	**24**	2	7,78	**3**
1	30,48	**21**	2	9,63	**8**
1	45,86	**23**	2	8,90	**6**

Test U Manna-Whitneya

Sumujemy rangi w każdej grupie.

Grupa I	Stężenie	Ranga	Grupa II	Stężenie	Ranga
1	14,34	**14**	2	5,33	**1**
1	20,33	**17**	2	22,50	**19**
1	18,79	**15**	2	11,74	**9**
1	8,22	**4**	2	7,39	**2**
1	31,50	**22**	2	12,34	**12**
1	12,08	**10**	2	13,22	**13**
1	22,00	**18**	2	8,53	**5**
1	9,22	**7**	2	22,80	**20**
1	19,50	**16**	2	12,70	**11**
1	78,89	**24**	2	7,78	**3**
1	30,48	**21**	2	9,63	**8**
1	45,86	**23**	2	8,90	**6**
		R_1=191			**R_2=109**

Test U Manna-Whitneya

Obliczamy wartość funkcji testowej U:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U = 12 \cdot 12 + \frac{12(12+1)}{2} - 191 = 31$$

Wartość odczytana z tablic wartości krytycznych testu U Manna-Whitneya dla poziomu istotności p=0,05 wynosi

$$U_p(n_1, n_2) = 37$$

Wnioskowanie

Jeżeli wartość funkcji testowej jest mniejsza od krytycznej, to hipotezę zerową odrzucamy.

Test U Manna-Whitneya

Było:

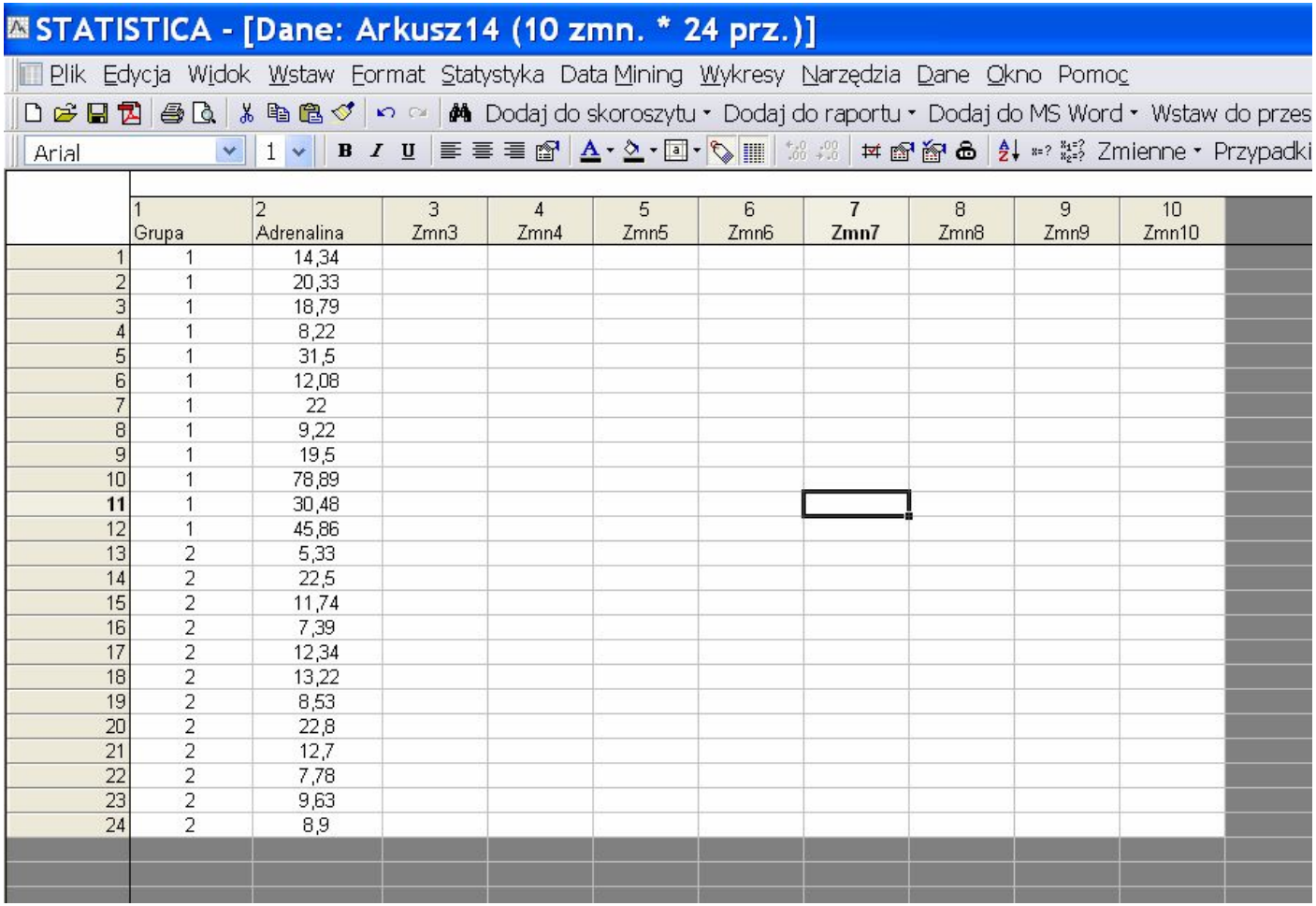
Hipoteza zerowa

próby pochodzą z tej samej populacji (z populacji o równych medianach)

Odrzuciliśmy H_0, zatem próby nie pochodzą z tej samej populacji – różnica między stężeniami (wyrażonymi medianami) jest istotna statystycznie.

Obliczenia w pakiecie *STATISTICA*

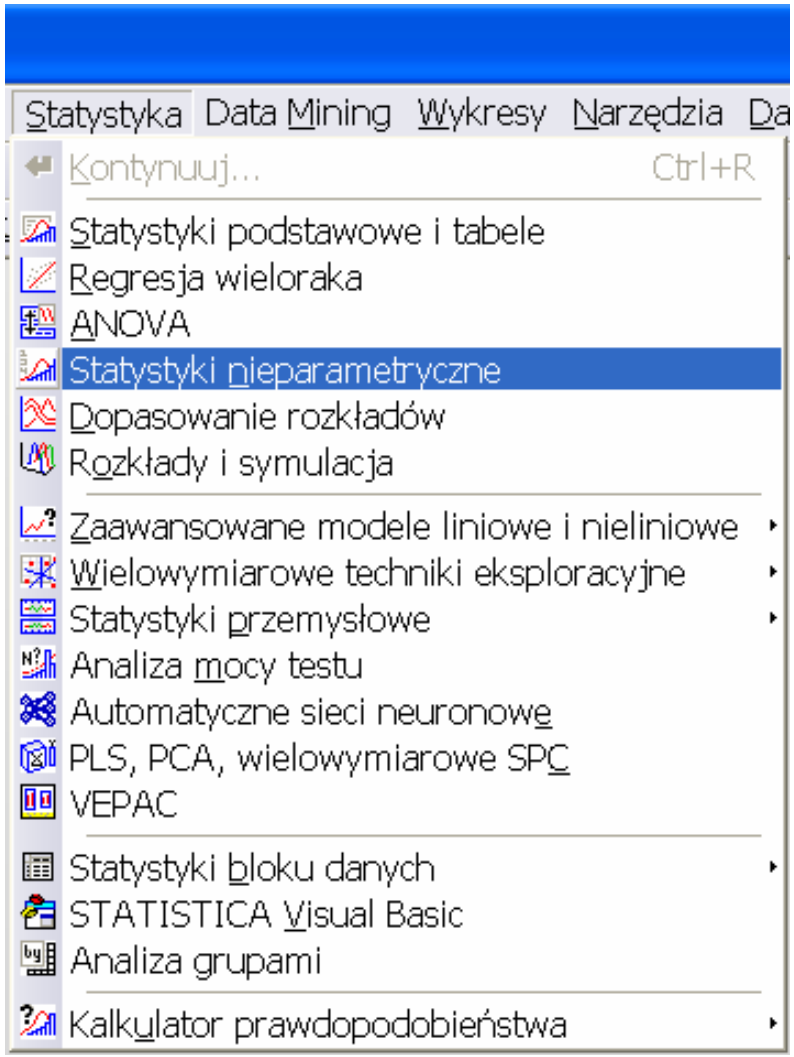
Dane rozmieszczone w dwóch kolumnach: w jednej mierzone wartości, a w drugiej identyfikatory grup (zmienna grupująca).



	1 Grupa	2 Adrenalina	3 Zmn3	4 Zmn4	5 Zmn5	6 Zmn6	7 Zmn7	8 Zmn8	9 Zmn9	10 Zmn10
1	1	14,34								
2	1	20,33								
3	1	18,79								
4	1	8,22								
5	1	31,5								
6	1	12,08								
7	1	22								
8	1	9,22								
9	1	19,5								
10	1	78,89								
11	1	30,48								
12	1	45,86								
13	2	5,33								
14	2	22,5								
15	2	11,74								
16	2	7,39								
17	2	12,34								
18	2	13,22								
19	2	8,53								
20	2	22,8								
21	2	12,7								
22	2	7,78								
23	2	9,63								
24	2	8,9								

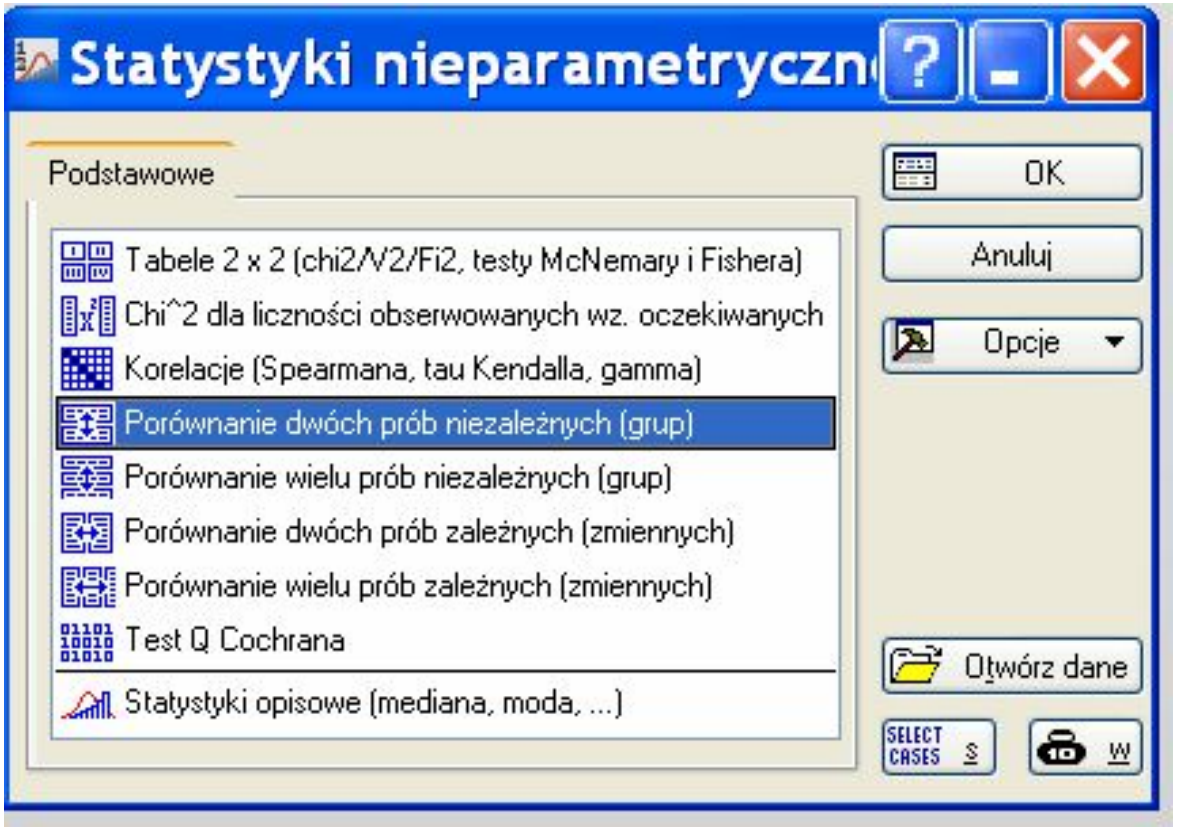
Obliczenia w pakiecie *STATISTICA*

Wybieramy z menu **Statystyka/Statystyki nieparametryczne**.



Obliczenia w pakiecie *STATISTICA*

W okienku **Statystyki nieparametryczne** wybieramy **Porównanie dwóch prób niezależnych (grup)**, **OK**.

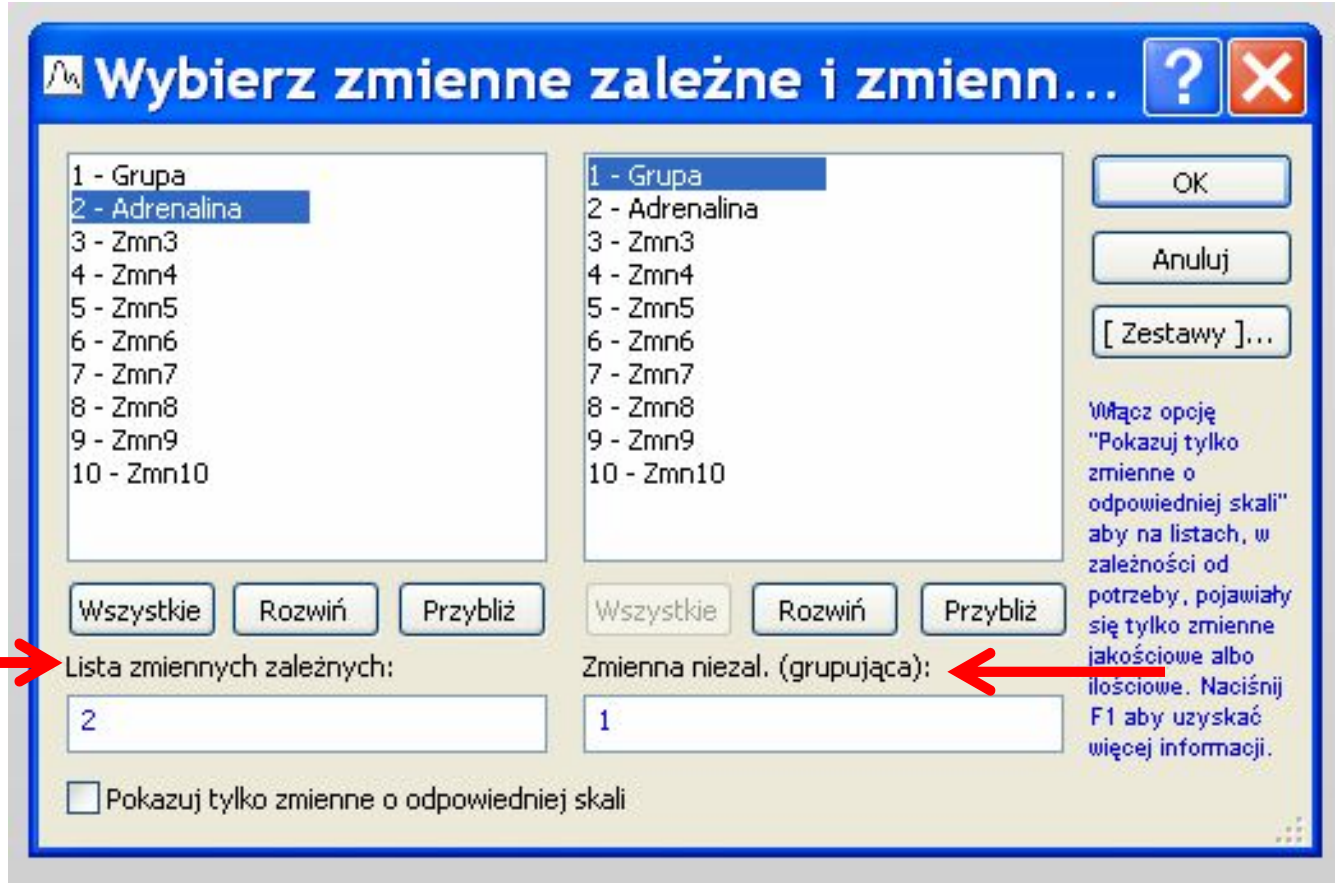


Obliczenia w pakiecie *STATISTICA*

W okienku **Porównanie dwóch grup** wybieramy **Test U Manna –Whitneya.**

Obliczenia w pakiecie *STATISTICA*

W okienku **Wybierz zmienne...** zaznaczamy na liście zmiennych zależnych **Adrenalina**, a na liście zmiennych grupujących **GRUPA**, **OK**.



Obliczenia w pakiecie *STATISTICA*

Z arkusza z wynikami odczytujemy ...

Test U Manna-Whitneya (z poprawką na ciągłość) (Arkusz14)

Test U Manna-Whitneya (z poprawką na ciągłość) (Arkusz14)
Względem zmiennej: Grupa
Zaznaczone wyniki są istotne z p <,05000

Zmienna	Sum.rang Grupa 1	Sum.rang Grupa 2	U	Z	p	Z popraw.	p	N ważn. Grupa 1	N ważn. Grupa 2	2*1str. dokł. p
Adrenalina	191,0000	109,0000	31,00000	2,338269	0,019374	2,338269	0,019374	12	12	0,017271

- w komórce U - wartość funkcji testowej dla prób o małych liczebnościach (poniżej 20)
- w komórce Z - wartość funkcji testowej dla prób o dużych liczebnościach (obie próby powyżej 20)
- w komórce Z popraw. - wartość funkcji testowej dla dla danych, w których występują rangi wiązane
- w komórce p – poziom istotności wyznaczony dla otrzymanej wartości testu

O seriach ...

Sikora bogatka

Sikora modra

O seriach ...

Przykład

Na pewnym obszarze pary lęgowe sikory bogatki (B) i sikory modrej (M) rozpoczęły składanie jaj w następującej kolejności:

B B M M B M M B B M B M B M B B

Seria to układ identycznych elementów leżących obok siebie, tu jest 11 serii.

Taki obraz nie świadczy o tym, żeby jeden z gatunków zaczynał lęgi szybciej.

O seriach ...

Przykład cd.

Inna możliwość:

B B B B B B B B M B M M M M M M

Tu są 4 serie i można sądzić, że bogatka zaczęła lęgi wcześniej.

W testach serii mała liczba serii świadczy o tym, że próby różnią się.

Test serii stosuje się do danych zapisanych w skali porządkowej.

Test serii Walda-Wolfowitza

Przykład

Na 10 poletkach w borze (B) i 9 w dąbrowie (D) odłowiono pająki krzyżaki w następujących liczbach

B: 37, 30, 45, 52, 22, 35, 27, 40, 47, 32

D: 48, 57, 31, 53, 51, 64, 44, 61

Czy można przyjąć, że poletka w tych dwóch typach lasu nie różnią się rozkładami liczb odłowionych w nich pająków?

Test serii Walda-Wolfowitza

Przykład cd.

Najpierw dane porządkujemy dla obu prób łącznie ...

B: 22, 27, 30, 32, 35, 37, 40, 45, 47, 52

D: 31, 44, 48, 51, 53, 57, 61, 64

B:	22	27	30		32	35	37	40		45	47			52				
D:				31					44			48	51		53	57	61	64

Test serii Walda-Wolfowitza

Przykład cd.

Następnie nadajemy im rangi:

B:	22	27	30		32	35	37	40		45	47			52				
	1.	**2.**	**3.**		**5.**	**6.**	**7.**	**8.**		**10.**	**11.**			**14.**				
D:				31					44			48	51		53	57	61	64
				4.					**9.**			**12.**	**13.**		**15.**	**16.**	**17.**	**18.**

Wartość funkcji testowej jest równa liczbie serii r. Tu r = 8. Wartość krytyczną odczytujemy z tablic dla poziomu istotności 0,05. Liczebność dla boru n_1=10, dla dąbrowy n_2=8. Wartość krytyczna wynosi $r_{0,05}$=6.

Wnioskowanie

Hipotezę zerową odrzucamy, jeśli liczba serii jest mniejsza od wartości krytycznej.

Test serii Walda-Wolfowitza

Pytanie postawione w tym przykładzie:

Czy można przyjąć, że poletka w tych dwóch typach lasu nie różnią się rozkładami liczb odłowionych w nich pająków?

Oznacza H_0: poletka w dwóch typach lasu nie różnią się istotnie.

Hipotezy zerowej nie można odrzucić, zatem nie stwierdziliśmy różnicy między poletkami w borze i dąbrowie.